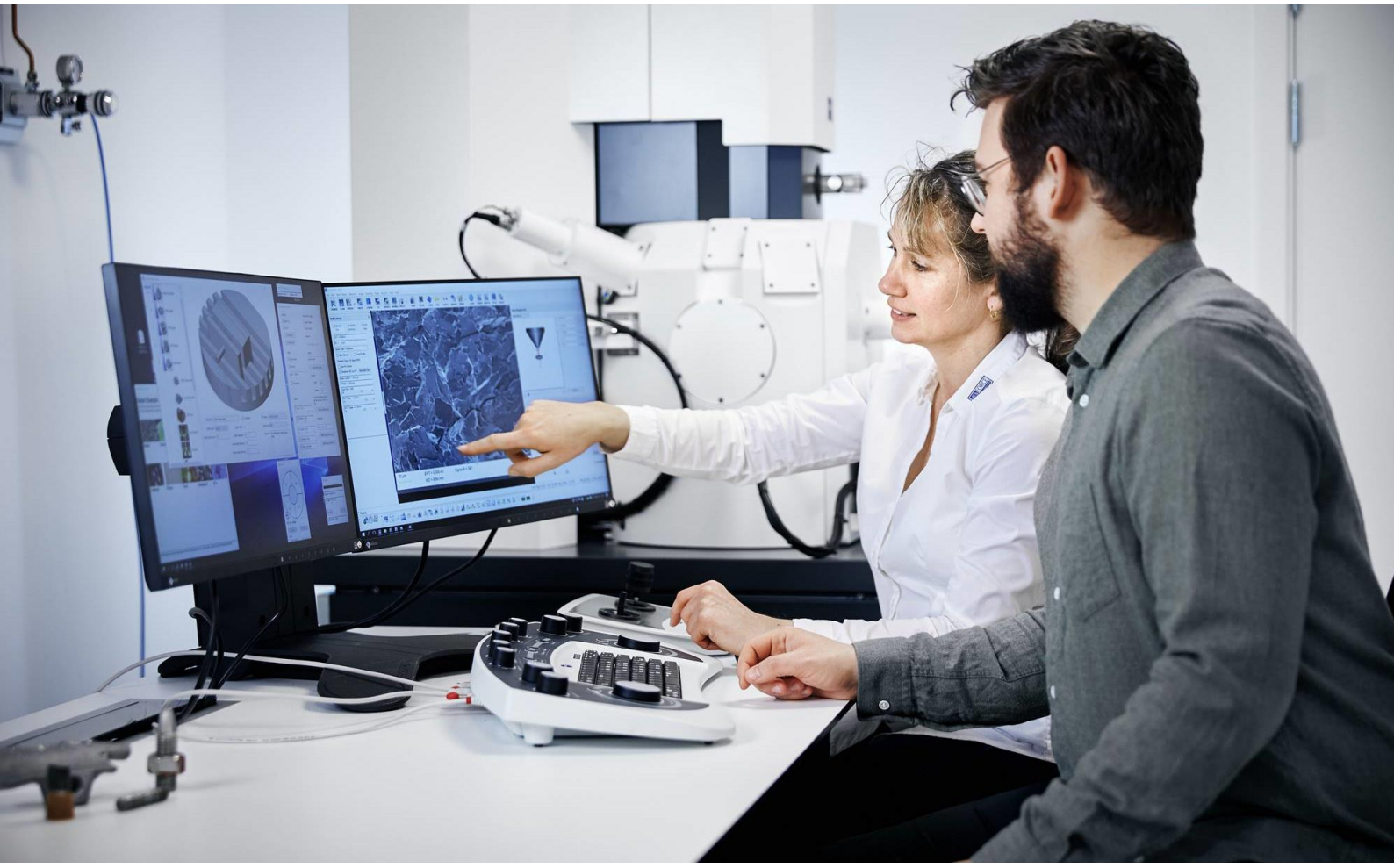
IoT Data Validation: what is data validation and why validate IoT data?



Anders P. Mynster, *Head of Department,*
IoT, Data & Services Innovation, FORCE Technology (apm@force.dk)

Anders J. Madsen, *Specialist,*
IoT, Data & Services Innovation, FORCE Technology (ajm@force.dk)

Mads Johansen, *Specialist,*
IoT, Data & Services Innovation, FORCE Technology (majh@force.dk)

Ashok Debnath, *Senior Specialist,*
IoT, Data & Services Innovation, FORCE Technology (asde@force.dk)

Table of Contents

Abstract

IoT is one of the key technologies in today's digital revolution. Connected devices, coupled with advances in AI, big data, and analytics, are driving new use cases across industries where IoT is helping enhance the end-user experience, create new business models, and improve operations and business planning. IoT infrastructure allows to capture of meaningful data and further enables innovative business models and opportunities. Enterprises are building variously connected and smart IoT applications to help improve efficiencies and deliver customer value across the full range of consumer and business-to-business scenarios.

IoT Technologies offer new opportunities in many areas like Smart Home, Wearable's device, Connected Automobile, Healthcare, Pumping Solution. Also, IoT helps in solving Smart Cities and societal challenges such as using data to manage traffic, parking, water supplies, efficient lighting, waste collection and disposal system, cut pollution and make better use of infrastructure.

Denmark has a solid foundation for developing IoT capabilities and using sensor data to drive future innovation. Danish companies and public service providers are constantly using data and digital technology such as IoT, AI, and Machine Learning to deliver services by responding to the Danish citizens' needs.

Considering the unique characteristic of IoT solutions, FORCE Technology is joining hands with Alexandra Institute to enable robust end-to-end data validation strategies for testing IoT enabled applications that will ensure higher stability, scalability, interoperability and provide rich user experiences.

The aim of project is to help organisations ranging from companies who just started using IoT technology to collect some data to companies which are already mature in using IoT technology and generates enormous amounts of data from different devices. And this IoT Data validation Solution will be designed to help manufacturers/companies modernize their operations and gain following benefits:

1) Create or increase trust in the IoT-systems
2) Fulfil requirements for operations compliance
3) Reduce the time spent on correcting errors in the IoT-system
4) Quickly react to faulty system before data is published further - before it reaches many users
5) Increase efficiency of the resources spent on the IoT-system
6) Detect fraud by users or vandalism to publicly available IoT-systems

In this whitepaper, we will discuss major challenges in IoT data validation, the Data Validation Stages, and lastly the Values & Services which will be enabled from basic validation to advanced validation. We will also discuss the integration and use of modern tools like AI, Advance statistics, Machine learning for data validation services such as Schema Validation, Data Visualization, Anomaly detection, Traceability, etc. to cover different sizes of companies.

1. Introduction

1.1. Background of IoT Data Validation Solution

As IoT systems become increasingly integrated into our day-to-day lives, it becomes essential to validate data generated from IoT systems to ensure that they are secure, efficient, perform their intended tasks, and comply with government regulations and industry standards. A typical IoT solution retrieves data from an array of heterogeneous sensors provided by a variety of vendors. In the industry, collectively need a method to spot any fault in the sensors, even before the service based on the data reaches the customers and end-users.

In the last decade, we have witnessed an exponential growth of IoT and today IoT has become one of the major contributors in large-scale systems. So naturally, there will wear and tear of the devices/sensors over the period and identification of these issues needs to be done intrinsically by analyzing the captured data. Also, the efficacy of the devices is important, and to manage that we need to analyze the captured data. IoT is also challenged to address security, privacy concerns, and network issues as they directly impact the reliability and accuracy of data. Thus, data validation for IoT data goes beyond just data cleaning, aggregation, and transformation, and shifts more towards intelligence and machine learning-based methods for data abstraction and predictive methods.

Data is becoming increasingly valuable in the industry due to the importance of data products such APIs, dashboards, benchmarks, and report creations. International Data Corporation (IDC) predicts that *"**every connected person in the world will have at least one digital data interaction every 18 seconds — likely from one of the billions of IoT devices, which are expected to generate over 175 ZB of data in 2025***" [1]. The role data plays in the decision-making process and the development of machine learning and deep learning models makes it even more important. Therefore, all processes associated with data ranging from data generation to data reception need to be monitored. Fault detection, reporting, anomaly detection and mitigating the effect of faults are complex but inevitable while building efficient data products.

IoT systems are extremely complex to validate, as they combine hardware devices and sensors, gateways, network components, internet, cloud, software applications and many more components. Moreover, the real-time features of IoT systems having characteristics of volume, velocity, and variability adds to the validation complexities. Also, companies must account for changing market conditions, and diverse product specifications. Together, these challenges make IoT testing time-consuming and complex.

IoT Data validation is an essential part of any data handling task, be it in the field of collecting information, analyzing data, or preparing to present data to stakeholders. If sensor data is inaccurate from the beginning, aggregated values or derived results will also be inaccurate. IoT data validation starts with a data source that generates data and ends at a destination that receives the processed data. Components in the data validation are capable of automating processes involved in extracting, transforming, combining, validating, and loading data. Moreover, data validations eliminate errors and accelerate the end-to-end data processes which in turn reduces the latency in the development of data products.

As the IoT ecosystem continues to grow, each domain (including Agriculture Production, Cities, Traffic, Infrastructure, Urban Assets, Manufacturing, Transportation, Home & Appliance, and other Domain) both generate and consume data. And data generated from different various domains in the IoT ecosystem leads to different applications such as intelligent e-health, smart meter, smart home, smart healthcare, logistics management, warehouse management, and many more. The possibility of erroneous, inaccurate, or inconsistent data is very high in most IoT deployments because they are based on heterogeneous sensor types. When the data is transformed from different data sensors or devices, several factors are affecting data quality. It may be affected by sensor fault, or environmental factors such as embedded in the concrete or attached on underground pipes or in the splash zone of the waves at sea, or they are simply installed outdoors, receiving all the harshness of the environment, also during data transfer and pre-processing, network outages may impact data quality, and also factors such as privacy preservation processing affect data quality. This in turn affects the applications built from such data. So, it is important to evaluate data before data is further used or published to make decision.

1.2. IoT data quality properties

One of the key aspects of IoT data validation is to align key properties that the validation should address. This is associated with the properties of the IoT data, necessary to build robust solutions, reliable data driven solutions, stability of the systems and in general – trust in the function of the IoT system. Each data flow using data from IoT devices should, based on the criticality of the operation and accuracy of the service the system is providing, fulfil certain data quality properties. Based on the current experiences of the Nordic IoT centre, literature studies and investigation of current state of the art platforms, these are as follows:

Figure 1 - Data quality properties

The objective of this whitepaper is to understand the existing IoT data validation as well as the challenges experienced at the case company and to develop a conceptual model of the robust data validation. Based on the study objectives, we will identify challenges related to IoT data validation that practitioners in the case company experience. We will also try to understand importance of IoT data validation Data Quality properties and its importance formulated and answer following questions:

Trustworthiness:
How to make data is trustworthiness on data generated from IoT Device? Why Trustworthiness is important for IoT Data validation? How Do You Ensure Trust in IoT?

IoT is still emerging technologies, and it consist of various IoT stacks. In IoT, businesses may be starting with a sensor that could be in the facility or out in the field, remotely operating. Data is collected, brought through a gateway, and then moved through a series of servers and applications before it finally comes to rest in the middle of the data-centre, where it can be analysed and acted upon. Data may be either forged in different intermediate layers when it traverses, or data is read incorrectly by IoT device from the sensor. So, in general there are two types of trust per-spectives: User Trust, and System Trust.

User Trust refers to “subjective expectation an entity has about another’s future behaviour”, While System Trust refers to “the expectation that a device or system will faithfully behave in a particular manner to fulfil its intended purpose”, few examples:

- Smart thermometers which measure Blood Pressure or Temperature of body allows data to be shared with a doctor or family caregiver for a second opinion and can collect logs that track health and symptoms over time. Thus, it needs to show precise data so User can trust the behaviour.
- Trust in healthcare-related information provided by variables can lead to critical decisions. For example, a user may incorrectly decide to not visit a doctor based on the vital diagnostics check from their device.
- If I delegate access to my home sensor information to my power utility, what can they do with the information and how is it protected?
- Smart wearable device there are several potential risks that the users foresee while using the wearables, for example the accuracy of fitness tracking, loss of sensitive data captured by the devices thereby violating privacy or even their reliability and overall usability.

Fulfil Compliances:
How to ensure that data is fulfilling general and standard compliance? Do your devices fulfil GDPR and other compliances?

Several standards have been developed for IoT connectivity, some of which address connectivity of low-power devices such as home security systems or Wi-Fi-enabled devices or Bluetooth or to the internet. IoT product follows standard compliance means that the people, processes, and technologies that make up an integrated and deployed IoT system are compliant with some set of regulations or best practices.

For example, an individual may be more inclined to use a Smart Wearable system for health monitoring purposes if it was FDA approved. The regulations such as FDA approval or clearance are structural assurances for the wearable system. These structural assurances strengthen trust of users in wearable systems and institutions associated (such as manufacturer, app-makers, etc.)

An additional dimension of the compliance in relation to IoT, is that often the IoT systems themselves are used to document compliance, such as in cold chain monitoring for documentation of correct temperatures during transport of goods. Hence there are multiple aspects of compliance for IoT systems

- Compliance regarding technical standards – often associated with interoperability and reliability
- Compliance regulatory requirements – often associated with safety
- Compliance regarding the use and origin of the data

Interoperability:
How to ensure interoperability between device, platform? How to achieve interoperability from the data generated using different device and platform?

With the evolution of IoT, there are many smart objects found in the physical world, which are interconnected and communicate through the existing Internet infrastructure. However, each solution provides its own IoT infrastructure, devices, APIs, and data formats leading to interoperability issues. So cross-platform interoperability between things and data in this scenario enables interoperability across separate IoT platforms specific to one vertical domain such as smart home, smart healthcare, smart garden, etc.

For example, a user who has health problems uses an IoT cross-platform application every day to help him with his everyday tasks. The IoT application connects to the user’s smart health platform of wearable sensors to continuously monitor his health conditions (heart rate, fall situation, and glucose level) and in an emergency, locates him and sends an ambulance.

Integrity:
How to ensure data integrity means ensuring that data is complete, original, consistent, and accurate?

In the context of IoT, integrity looks after the data contained within the device or system while availability covers accessibility of the data from the device. Any breach of data integrity will mean that an IoT device cannot operate correctly but it also potentially exposes the device to being exploited and become a compromised platform from which other attacks can be launched. Integrity checks also needs to be used for data that is being processed to ensure that the data and its flow can be trusted.

For example, invalid temperature data can cause the control unit of an industrial establishment to turn on/off the cooling system arbitrarily, which can cause significant damage to the machinery and may even result in personnel injury.

Thus, data is the life blood of IoT operations and it is critical that its integrity is robust. All parties involved must ensure their data has not been manipulated or tampered with while at-rest, in-transit, or in-use.

Accuracy:

How accurate data generated from IoT device? And to use data to generate billing to your customer?

Accuracy is how close the measured value is to the true value. The values that have been accumulated from across the network of IoT devices accurately reflect what was produced at each device. Further data generated from IoT device can also be used to generate billing to the customer like Smart Meter for Water, Electricity utilities. In specific applications accuracy also needs to be documented, such as through traceable or accredited calibrations.

Moreover, the characteristic of IoT data is very heterogeneous in the sense that different IoT devices produce different types and format of data. This means that even the conversions from one format to another must be validated. In addition, there is inaccuracy in the sensed data, there are scalability problems in the data because the devices may go offline and finally data semantics are different from one device to another. This can potentially lead to a lack of accuracy on a high-level dashboard displaying the sum of thousands of sensors, but if even 5 devices report wrongly formatted data, it can heavily impact the accuracy of the overall result. Therefore, data management systems must be able to cater for such heterogeneity in IoT Data.

Of particular interest for IoT systems is also the accuracy of time stamps associated with the data. This is due to the fact that many IoT systems need to be low power, which reduces the possibility to keep a stable clock, and to synchronize time stamps with the network. A data value might be collected with a very high accuracy, but if the true value, has deviated during the time difference between the device and the network, the measurement becomes inaccurate. For additional information, see timeliness.

For example, Smart Meter calculate consumption of water and electricity usage and further generate accurate billing for the period on desired time. So, it is important that data generated for the usage is accurate and further billing should be done on the active usage of consumption and should not calculate on inactive usage.

Consistency:

Are the values logged with IoT device is consistent with the context of values were produced by each device?

Consistency of data refers to the correlation among a series of data points that are reported by the same sensor over a short period of time. IoT devices are designed to take a sensor reading on a schedule and record and report the result. This process is automated, meaning that the schedule is consistent and results in large amounts of data being captured. Inconsistent data leads to anecdotal reporting, while consistent data over time develops a pattern and is much more compelling.

For example, a temperature sensor is not expected to report very different temperature readings or wild swings in temperature within a few seconds. Or a geolocation sensor is not expected to report locations that are many kilometres apart within few seconds. If such a case happens, it is likely that the sensor is faulty. A sensor might report readings that are consistent but incorrect if its calibration is off. A sensor that regularly reports data that is inconsistent with readings from other sensors might automatically be reported for health check.

Completeness:

Are there are any gaps in the series of reported events or sensor values that should have been captured?

Completeness of data refers to whether all supporting data points are available; for example, raw data that contains all values which supports a predicted event is available in the system or time series data that does not have missing data points. Quality data is when meta-data points are available, which describes how the recorded data can be traced back to a particular sensor at a particular point of time. Where relevant, completeness can refer to the ability to combine and correlate data from multiple sensors or even from other information systems.

It is essential to understand that completeness is related to the purpose of using the data, i.e., the IoT-based service being delivered to the users of the IoT system. Some IoT services can be completely represented by a single measured data value recorded on an hourly basis, and some need streams from thousands of sensors updated every second.

Timeliness:
Are the values being captured within a reasonable time frame?

As so much of IoT data is real-time data or near real-time data, determining whether the sensor data or derived data arrived on time at the required point in the network becomes critical. It is desirable to be able to act on the data in time to prevent or pre-empt incidents. It is important to correlate readings from multiple sensors, timeliness might include the ability to synchronize the data from those sensors.

If much of the data is streamed and coming from a wide variety of devices, are there monitoring points to ensure that the collective data set is synchronized?

Reliability:
How reliable data collected from IoT Device?

Reliability is must when we have mission critical applications. For reliability, the measurements from the sensors must be accurate and repeatable, over a given lifetime of the sensor. Accuracy of the reading refers to the required precision of the measurement that must be achieved by the sensor. Repeatability refers to whether the sensor will produce the same measurement with the same accuracy when put in the same scenario. For example, if the geolocation sensor is brought back to the same street corner, the sensor needs to produce the same location readings to the precision required.

1.2.1. Objectives of IoT data validation

In addition to the above properties of the data, the data validation must also improve overall productivity and detect anomalies in general using data validation. Further we will try to answer question of below question:

Efficiency:
How to improve overall business efficiency and productive to take timely data driven decision making using of data validation process?

IoT has dramatically changed the way businesses operate and utilize technology. Improving connectivity to an unprecedented extent, IoT solutions streamline many functions for greater efficiency, reduce costs, and provide useful data from which a company can conclude which of their practices to keep and which to change.

The huge volumes of new data from IoT sensors and devices simply adds to the massive pool of data and using that data to inform strategic and operational decision making. Strategic decision-making is where the senior leadership team identifies the critical questions it needs answering. Further business entities have better understanding of customers which will allow them to use data to improve product and services.

Detect Anomalies:
How to automatically detect anomalies in the new data through a programmatic data validation process?

The IoT systems are being used in many diverse applications that are part of our life and is growing to become the global digital nervous systems. This will result in a potential change in the way we work, learn, innovate, live, and entertain. The heterogeneous smart sensors within the Internet of Things are indispensable parts, which capture the

raw data from the physical world by being the first port of contact. However, during course of time sensors are prone to failure, malfunction, rapid attrition, malicious attacks, theft, and tampering. All these conditions cause the sensors within the IoT to produce unusual and erroneous readings, often known as outliers.

Sensor faults and outlier detection is very crucial in the IoT to detect the high probability of erroneous reading or data corruption, thereby ensuring the quality of the data collected by sensors. The data collected by sensors are initially pre-processed to be transformed into information and when Artificially Intelligent (AI), Machine Learning (ML) models are further used by the IoT, the information is further processed into applications and processes. Any faulty, erroneous, corrupted sensor readings corrupt the trained models, which thereby produces abnormal processes or outliers that are significantly distinct from the normal behavioural processes of a system.

1.3. Who needs IoT Data Validation Solutions?

With today's IoT eco-system offering rapid and exponential changes in data volume, variety, and velocity, organizations may encounter errors and redundancies in the data collected from various device, sensor(s), and actuator(s). To address modern-day data validation challenges with an open source-based scalable solution FORCE together with Alexandra Institute is creating generic IoT Data Validation Centre, a best-in-class data automation software solution along with consultancy offer that will help streamline and accelerate the validation of data integration and data analytics platforms. It will be extension of Trouble shooting service introduce in Year 2020.

With these IoT Data Validation Solution, each company will take charge of records and automates validation of data regularization and accuracy. It will bring together a solution based on Cloud for a comprehensive technical approach to data validation between source and target machine/database/files and further do data quality check (duplicate records, missing records), data reconciliation checks, and data validation checks (duplicate records, incorrect pattern, missing records). IoT Data Validation Solution will enable your data quality assurance to achieve complete data coverage during different phases, reduce your time to market by identifying anomalies in data earlier in the life cycle, and thus lead to better customer satisfaction.

It is a solution that will enable any organization to strengthen their quality assurance processes to make sure that the data generated from IoT ecosystem is consistent, relevant, and timely. Built on open and adaptive architectural principles, solution will integrate with existing infrastructure, enabling access for teams across organization for better collaboration.

IoT Data Validation System will be modern solution-based on Cloud, AI and ML offering that automates the data validation process. Built on open and adaptive architectural principles, the solution integrates with your existing infrastructure to enable better collaboration. Some of the categories of sectors which can gain benefits of IoT data validation system:

- Government Sectors
- Companies that started their IoT journey
- Mature IoT companies which generate enormous amount of data
- Start-ups, SMEs
- IoT Enabled Consumer Segments Companies in areas: Home, Health, Mobility, Retail, Agriculture
- IoT Enabled Business Segments Companies in areas: Telecom, Enterprise, Manufacturing, Smart Cities, Public Sector, Health, Mobility, Retail, Agriculture, Utilities, Retail, Construction & Building Materials.
- Nordic IoT companies

Table 1 - Examples for Different Danish Industries where IoT Data Validation will be useful

Security	**Healthcare**	**Mobility**
- Predictive policing - Smart surveillance - Emergency response optimization - Disaster early-warning systems - Personal alert applications - Home security systems - Data-driven building inspections - Crowd management	- Remote patient monitoring - Lifestyle wearables - First aid alerts - Real-time air quality information - Infectious disease surveillance - Integrated patient flow management systems	- Real-time public transit information - Intelligent traffic signals - Smart parking - E-hailing (private and pooled) - Car sharing - Bike-sharing - Integrated multimodal information Real-time road navigation - Parcel load pooling - Smart parcel lockers
Energy	**Water**	**Waste**
- Building automation systems - Home energy automation systems - Home energy consumption tracking - Smart streetlights	- Water consumption tracking - Leakage detection and control - Smart irrigation - Water quality monitoring	- Digital tracking and payment for waste disposal - Optimization of waste collection routes

1.4. Features & Benefits of IoT Data Validation Solution

Below is the list of IoT Data Validation highlighted features also we are in process of adding sophisticated feature:

- Anomaly detection: In terms of equipment usage and function, production line, medical data etc.
- Visualization: Dashboards, Historical Correlation, Historical usage analysis
- Advisor service: Usage pattern detection and decision support system
- Predictive maintenance: usage prediction, demand-supply prediction
- Analytics to extract context from raw sensor data
- Supply chain analytics enhanced with real-time data
- Real-time alerts
- Web based queries and reports
- Advanced Statistics
- Basic Validation to Advance AI based Validation

Some of the overall benefits using IoT data validation as follows:

- Reduced support cost, lower breakdowns, improved operational efficiency
- Optimal scheduling of production lines
- Eliminate waste, improve efficiency, and optimize operations. Have trustworthy real-time data to improve productivity on the fly.
- Anomaly Detection by integrating machine learning to analyse current conditions to identify deviations from normal operating behaviour and correct potential errors in the data.
- Easy to reconfigure the underlying cloud components based on customer need.
- Creation of newer business models based on condition and usage.

2. The IoT Data Validation activities

IoT Data Validation Centre helps organizations validate data generated from different IoT system and ensures data quality to maintain accurate data flow across systems. It addresses the data challenges that arise because of security threats, lack of compliance adherence and data unavailability using solutions that are both on-premise as well as cloud native. IoT Data Validation Centre service offers early validation through Machine Learning, Data pipeline, and Analytics and Data Visualization. It also enables seamless integration into new data architecture to address challenges which existing traditional tools and data automation solutions cannot address. With exponential changes in volume, variety, and velocity of data, IoT Data Validation Solution will enable your organization to strengthen their quality assurance processes to make sure that the data shared is consistent, relevant, and timely.

In this whitepaper, we present processes such as “validation rules”, “validation constraints”, or “check routines”, which check for completeness, consistency, accuracy, reliability of data provided as input to the system. In other words, checking for correctness and meaningfulness of the data into the system. The rules may be implemented through an automated schema validation, data dictionary, or by the inclusion of explicit application program for data validation.

It is important to highlight that data validation does not only necessarily address accuracy, and it is possible for data entry errors such as misspellings to be accepted as valid. However, data validation is a process of identifying inaccurate, inconsistent data generated from the system, therefore reducing the number of errors, and indirectly the time can be spent on identifying the last errors.

In evaluating the basics of data validation, generalizations can be made regarding the different kinds of validation according to their scope, complexity, and purpose. It includes:

Activity	**Description**
Format validation	Checks that the input data is in the right format. For example, a Temperature is in the form XX °C where X is any number.
Data type validation	Verifies that the individual characters provided through user input are consistent with the expected characters of one or more known primitive data types
Check digit	It is used to find out if a series of numbers has been keyed, transmitted and received correctly. There are many ways to produce check digits. For example, Bar code readers check digit
Uniqueness check	Checks that each value is unique. For example, Machine unique identifier
Range and constraint validation	Checks the data falls between an acceptable upper and lower value, within a set range or consistency with a test for evaluating a sequence of characters, such as one or more tests against regular expressions. For example, a Temperature of room will be in range between 4 °C to 30 °C.
Code and cross-reference validation	It includes operations to verify that data is consistent with one or more possibly external rules, requirements, or collections relevant to a particular organization, context or set of underlying assumptions.
Structured validation	It allows for the combination of other kinds of validation, along with more complex processing. Such complex processing may include the testing of conditional constraints for an entire complex data object or set of process operations within a system.
Consistency validation	Consistency validation ensures that data is logical. For example, the delivery date of an order can be prohibited from preceding its shipment date.

3. IoT Data Validation Process

In this whitepaper we will focusing on creating process of IoT Data validation standard practice or state of the arts ways to validate date generated from IoT ecosystem. End-to-end IoT validation will be built on open and adaptive architectural based which consist of following steps:

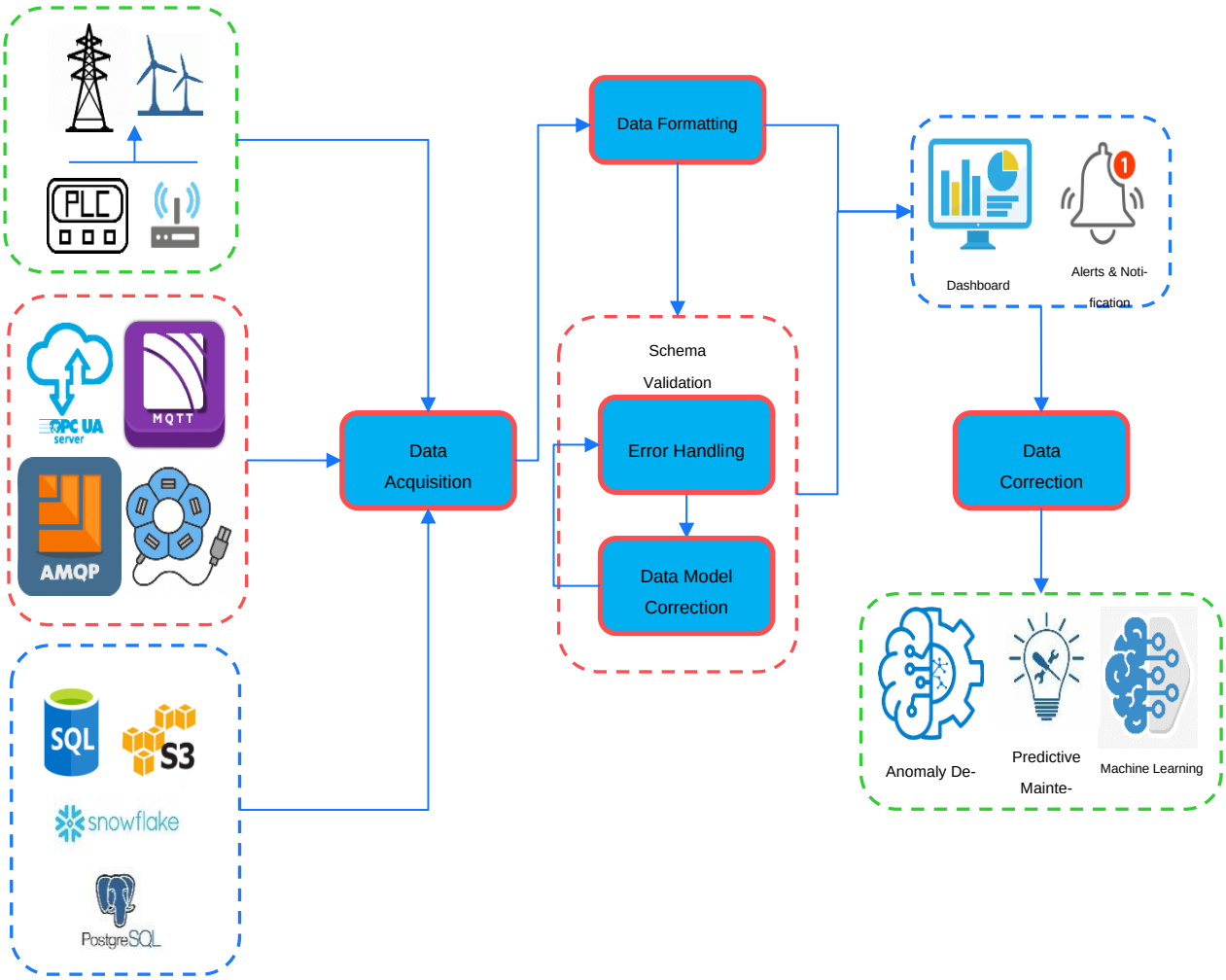


Figure 2 - Proposed - High level architecture for IoT Data Validation Process

Figure 3 - Proposed IoT Data Validation Process

3.1. Data Acquisition

During data acquisition step you can gather and collect data by using processes used to collect information. In the context of IoT there are numerous ways you can gather data. Data acquisition systems serve as a focal point in a system which allows to connect data source or device or products.

3.1.1. Connect to Data Source

During this stage you connect data source where data is located. It could be following: -

- Connect to Backend Database where database is located in VMs, On Premise, Cloud.
- Connect to Device, Hub or Edge using protocol like MQTT, UPC-UA, Message Queue such as Kafka
- Data can be connected via Gateways, APIs, File System (.CSV, JSON, XML etc.)
- User Interface

Data is collected from different source which has following properties: -

- When data is collected from a device located in customer network, it should be according to the legal agreement
- Different data sources generate data in different frequencies and formats
- Data collection mechanism itself should be capable to adjust with different intensities of data flow.

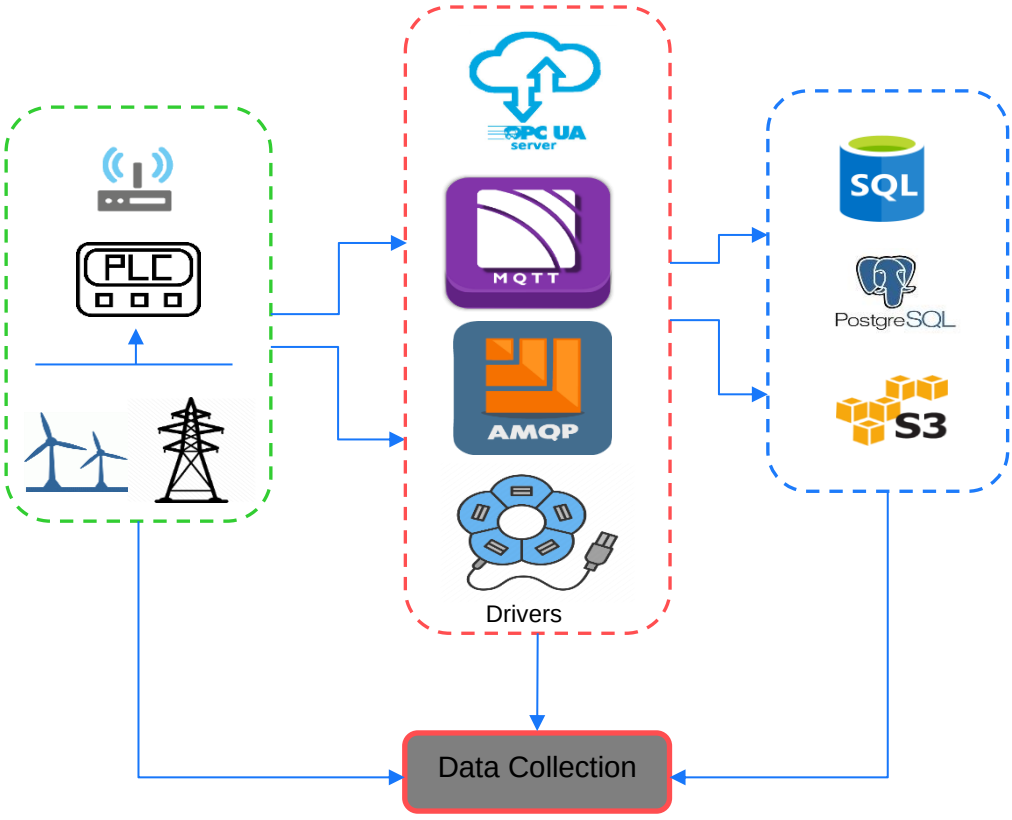


Figure 4 - Example of IoT Data Acquisition structure

Data collection agents can interact either with the nodes directly or with the device via protocols or connect to permanent data storage located in on-premise, cloud or VMs to collect the data. However, to ensure that the data collection agent has the right to collect data, it is authenticated with the help of access service.

3.1.2. Data Formatting

Data Formatting will ensure to make data conversion and converting into specific format for the further process of schema validation, data visualization and further do anomaly detection & troubleshooting. Data Formatting ensure consistency in naming conventions, default values, semantics, security while ensuring quality of the data.

3.2. Schema Validation

Picking a data model that is as standard as possible provides some valuable guarantees of interoperability, reusability, quality, but to live up to those promises, a formal validation is needed. One of the main technical approaches to data validation is through data schemas. They can be seen as spelling & grammar checker, but for computer data.

The primary goals of using data models are to:

- ensure interoperability and reusability.
- ensure that all required data objects are accurately represented, to avoid faulty reports and incorrect results.
- improve data quality and enable stakeholders to make data-driven decisions across several datasets.

3.2.1. Schemas for different communication layers

Basic standard formats: Especially when transferred from one system to another, data is often serialized using a basic standard format such as XML, JSON, CSV. This already ensures a standard way to parse the data, and provides a first sanity check, including lower-level aspects such as character encoding and Unicode validity. Compliance at this layer is assessed by a parser and not schemas, which intervene in the layers just above.

Specialised forms: Indeed, those basic standard formats are sometimes used in a specialised standardised form, for instance GeoJSON (JSON for geographical features), JSON-LD (JSON for linked data), or RDF/XML (XML for semantic data). Each of those specialisations can be validated through a combination of schemas and dedicated tools. There can be several layers of specialisation, for instance the W3C Web of Things (WoT) Thing Description builds on top of JSON-LD, which builds on top of JSON.

Ontologies: If not already decided by a specialised form, one or more ontologies can be picked to supplement a base format. Ontologies are special vocabulary for specific aspects or domains. For instance, how to express noise levels, or relations between individuals. The fragments of data using a given ontology can most of the time be validated against the official schema of that ontology. However, the official schema for an ontology might be for another serialisation than the one we have (e.g., RDF/XML instead of JSON-LD), in which case a transformation of either the schema or the data is needed (e.g., with XSLT, JSONata).

For a given piece of data, some parsing followed by one or more schema validation is thus typical.

3.2.2. Schemas and their specificity

Although a good start, it is not sufficient for a dataset to comply with a basic format (e.g., JSON or even JSON-LD) to be meaningfully valid. Indeed, some additional rules about the actual content should be incorporated. That means either using a specific data model, such as one from FIWARE Smart Data Models[1], or defining some custom/proprietary rules.

An example of specificity could be that the data must have a timestamp in a specific format, a sensor type from a list of know sensors, a geographical coordinate in a standard format, and a sensor value. That can be made more or less strict by allowing additional information or not, and by restricting the ranges of acceptable values.

3.2.3. Schemas and levels of sophistication

Most standard schemas are static and do not incorporate contextual information, such as based on the current time, day of week, season, information from previous observations, inter-documents constraints, or other business rules. However, such rules can be very valuable, and they are typically expressed through a distinct layer of validation, using another schema language (e.g., Schematron).

3.2.4. Error Handling

During the data transmission, its parsing, or its validation against a schema, different categories of errors will be detected. For instance:

- data losses (e.g., out of sequence messages)
- corrupt packets (e.g., checksum error, truncated message)
- data format checks (e.g., parsing, schema)
- validity checks (e.g., database constraints, business logic)

3.2.5. Data Model Correction

Data Model Correction is the process of correction of data model from the schema validation process. In Data Modelling stage we organize data description and understand constraints of data. The data model emphasizes on what data is needed and how it should be organized instead of what operations will be performed on data.

- Consistency checks
- File existence/missing values
- Uniqueness check
- Range Check
- Limit check

During this process we can correct Data model using both pre-defined and calculated values. (e.g., Fahrenheit to Celsius). Data Model Correction also performed adaptation to desired Data Model (e.g., remove unrecognized special character to fit data model)

3.3. Data Analytics

Data analytics refers to the process of examining datasets to draw conclusions about the information they contain. Data analytic techniques enable you to take raw data and uncover patterns to extract valuable insights from it. The main purpose of data analysis is to find meaning in data so that the derived knowledge can be used to make informed decisions.

3.3.1. Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. The goal of the data visualization is to communicate information clearly and efficiently to users by presenting them visually. So, data read from Data Model, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions. Data is represented in different visualization format such as:

- Charts & Diagram: Graphs, Tables, Histogram, bar chart, maps, infographics, line graph, scatter plot, heat map, line chart etc.
- Comparison of multiple parameters over time: Variance and averages
- Filter, Search and Sort options

- Saving data in different views or external file format like csv, excel, pdf etc.

3.3.2. Anomaly Detection

Anomaly detection is the identification of rare events, items, or observations which are suspicious because they differ significantly from standard behaviours or patterns. Anomalies in data are also called standard deviations, outliers, noise, novelties, and exceptions. Outliers are the data objects that stand out amongst other objects in the dataset and do not conform to the normal behaviour in a dataset. These anomalies might point to unusual network traffic, event detection in sensors network, or simply identify data for cleaning, before analysis.

We can perform Anomaly Detection on time-series based data and do both supervised and unsupervised anomaly detection using Machine Learning tools and perform average, variance and standard deviation to catch outliers.

Anomaly Detection will be also used to perform traceability and perform gap analysis (such as lost network connection, device malfunction, dead battery, calibration etc.). We can used various anomaly techniques such as:

- Simple Statistical Methods
- Density-Based Anomaly Detection
- Clustering-Based Anomaly Detection
- Support Vector Machine-Based Anomaly Detection
- Machine Learning-Based Approaches
    - Supervised Machine Learning for Anomaly Detection
    - Unsupervised Machine Learning for Anomaly Detection
    - Semi-Supervised Anomaly Detection

For additional information on methods and algorithms for Data validation, please see ”Algorithms for Validation of IoT Data, FORCE Technology, 2021”

Table 2: Identified list of tools which will be used for IoT Data Validation purpose.

Time Series Database	**Cloud Monitoring Tools**	**Anomaly Detection**
- InfluxDB - TimescaleDB **Relational Database** - Microsoft SQL Server - PostgreSQL **Search Engines Database** - Elasticsearch - Splunk **Key Value Stores Database** - Redis - Azure Cosmos DB **Document Store Database** - MongoDB - Azure Cosmos DB	- New Relic - Hyperic - SolarWinds - Retrace - Datadog - Dynatrace **Schema Validation Tools** - Google’s Structured Data Markup Helper and Data High-lighter - Schema Pro - GeoJSON Viewer & Validator - JSON Validator - XML Validator	- Datadog - Anodot - Elastic X-Pack - Splunk Enterprise **Machine Learning** - TensorFlow - Keras - Python - ElasticSearch **Stream/Data Processing** - Apache Kafka - Apache Spark

3.4. Technology Challenges in IoT

When you are ready to launch IoT based product/services, there are still some practical challenges you may encounter.

Device diversity and interoperability: Take an example where you are different types of sensor devices from different vendors. Also, as your product grows where you need to data exchange to other product/service supported by different series of sensor devices. As many vendors do not support any standards in their products, there are sure to be interoperability issues.

Integration of data from multiple sources: As you deploy an IoT application, you will get streams of data from different sources such as sensors, contextual data from mobile device information, and social network feeds and other web resources. It is important to note that the semantics of the data must be part of the data itself and not locked up within the application logic in different application silos.

Scale, data volume, and performance: Prepare your business to manage the scale, data volume, and velocity of IoT applications. As the number of users and devices scale, so will the amount of data that needs to be ingested, stored, and analysed. You will have a Big Data problem on your hands, and standard architectures and platforms may be inadequate. Also, where stringent real-time performance is required, network and application level latencies may be a problem.

Flexibility and evolution of applications: You will witness sensors and devices evolving with new and improved capabilities. This will result in creation of new analytics techniques and algorithms, and new use cases and business models. You will need to quickly develop apps with minimal effort. You will need ecosystems and platforms that enable and sustain this.

Data privacy: A good bit of data collected from devices will be sensitive personal data that must be protected from unauthorized access and used only for the specific purpose for which the user has allowed that data to be collected. Users must be provided with necessary tools that enable them to define the policies for sharing their personal data with authorized persons and applications.

3.4.1. Consequential Data validation challenges

- Lack of data accuracy, integrity and completeness while migrating from legacy sources
- Different kinds of data models, types and organization across legacy sources and packages
- Large volume of historic data that needs data analysis and conversion
- Complex data flow changes across the interfacing applications for integrated reporting and analytics

References

1. Coughlin, T. (2018, November 29). *175 Zettabytes By 2025*. Forbes. Retrieved in March 2021 from: https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/?sh=750a5d735459
2. All, I. F. (2020, June 4). The Importance of Multistage Validation to Successful IoT Solutions Development. IoT For All. https://www.IoTforall.com/the-importance-of-multistage-validation-to-successful-IoT-solutions-development
3. Business and Technology Trends - Smarter With Gartner. (n.d.). Copyright (C) 2021 Gartner, Inc. All Rights Reserved. https://www.gartner.com/smarterwithgartner/
4. Cities: a “cause of and solution to” climate change. (2019, September 23). UN News. https://news.un.org/en/story/2019/09/1046662
5. Coughlin, T. (2018, November 29). 175 Zettabytes By 2025. Forbes. https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/?sh=750a5d735459
6. Data Quality Dimension - an overview | ScienceDirect Topics. (n.d.). Data Quality Dimension - an Overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/computer-science/data-quality-dimension
7. Flender, S. (2019, February 11). Data is not the new oil - Towards Data Science. Medium. https://towardsdatascience.com/data-is-not-the-new-oil-bdb31f61bc2d
8. Kx. (2020, September 24). for Sensors: Data Validation, Estimation and Editing for Utilities and Industrial IoT. https://kx.com/blog/kx-for-sensors-data-validation-estimation-and-editing-for-utilities-and-industrial-IoT/
9. Ludwig, M. (2021, January 26). Why Companies Need Data Validation. RTInsights. https://www.rtinsights.com/how-to-deal-with-data-validation/
10. Mendis, A. (n.d.). Data Validation for Machine Learning. KDnuggets. https://www.kdnuggets.com/2020/01/data-validation-machine-learning.html
11. Sensor node data validation techniques for realtime IoT/WSN application. (2017, March 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/8166984
12. TensorFlow Data Validation: Checking and analyzing your data | TFX. (n.d.). TensorFlow. https://www.tensorflow.org/tfx/guide/tfdv#schema_based_example_validation
13. Transforming Data With Intelligence. (2016, February 19). Transforming Data With Intelligence. https://tdwi.org/articles/2016/02/19/six-validation-techniques-data-quality.aspx