

MADE White Paper

Empowering Production: Data Validation and Visualization for Strategic Excellence



Kasper Stens Honoré
M.Sc.
Technical Project Manager
FORCE Technology
ksho@forcetechnology.com
+45 4262 7042

Mads Johansen
M.Sc.
Specialist
FORCE Technology
majh@forcetechnology.com
+45 4262 7531

Ahmed Khan Leghari
PhD.
Senior Specialist
FORCE Technology
ahkl@forcetechnology.com
+45 4262 7160

November 2024

Table of content

- Introduction
- What is Data Validation, and Why is it Important
- What is Data Visualisation, and Why is it Important
- Data Processing: Node-RED
- Data Visualisation: PowerBI & Grafana
- High level architecture of a sample use case for a data-driven closed loop control application
- Conclusions and recommendations

This white paper sums up knowledge and learnings about data validation and visualization drawing on the results from relevant MADE projects with the manufacturing industry.

Introduction

In an era of rapid technological advancement, data-driven decisions have become integral to maintaining a competitive edge. This report delves into the crucial areas of data validation and visualization, emphasizing their impact on modern production environments. Through automation, AI, and powerful visualization tools, companies can harness data to enhance operational efficiency, minimize errors, and drive business growth. This report aims to provide insights into the importance of data quality, the use of tools like Node-RED for automation, and the benefits of visualization platforms like Power BI and Grafana for strategic decision-making.

What is Data Validation, and Why is it Important

Definition of Data Validation

Data validation encompasses a range of techniques and practices aimed at identifying data issues to ensure that data can be trusted and used effectively. Data validation can also be used to avoid introducing errors, and in some cases, the rectification of errors is included in the data validation process.

According to ISO 8000-2:2022, data quality is defined as the “degree to which a set of inherent characteristics of data fulfils requirements.” Furthermore, ISO 8000-2:2022 defines validation as “confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled.” In other words, since quality requirements can vary from case to case, data (quality) validation is not just about ensuring high-quality data but about ensuring data is fit-for-purpose, reliable, and trustworthy.

For example, data gathered from a simple home indoor climate sensor does not need to adhere to the same validation criteria as data gathered from sensors in a medical production facility, where there are strict requirements for traceability and accuracy.

To perform data validation, requirements for data quality must be defined. These need to be measurable, and different metrics should be created to address various aspects of data quality corresponding to the requirements and needs of the system and its stakeholders.

Importance of accurate and reliable data

Data validation is a crucial step in ensuring the reliability and quality of data used for making decisions and gaining insights through analysis. If data is not trustworthy, neither is the analysis nor the conclusions based on that analysis.

Highly accurate data is essential in manufacturing, where machines often work with very small tolerances. But reliable data is not only important for operating machines. It is also important for gathering business data, where data might not be collected with the same degree of accuracy, but with reliable business data, production planning and inventory management can be optimized, risks and waste can be reduced, and overall productivity can be increased.

Reliable data gathered from sensors on manufacturing machines can tailor maintenance to individual machines, increasing production quality. Similarly, if sensors are placed on finished products where relevant, more reliable services can be offered.

Consequences of poor data validation

There are several negative impacts of poor data quality, which result from poor data validation:

1. Wrong Conclusions and Decisions

If conclusions or business decisions, such as production planning, are based on poor data, this can lead to missed opportunities and other problematic situations as mentioned in the following points.

2. Increase in Operational Costs

In production environments, poor data validation could lead to increased operational costs. For example, a machine running under suboptimal conditions may consume more power, or a worn machine might not receive maintenance in time (predictive maintenance), leading to inefficiencies due to data not being monitored correctly.

3. Affecting Downstream Users

If a system presents poor-quality data to downstream users who rely on it for decision-making or if a system acts autonomously by adjusting parameters or actuators, it can negatively impact downstream users. In a production setting, this could lead to delays or quality issues, and customer deadlines may not be met.

4. Lost Revenue

Poor data can lead to lost revenue in several ways. For example, incorrect customer or sales data in a production company can result in missed opportunities or reduced profits, compounded by the issues mentioned in the other points.

5. Reputational Damage

When customers or downstream users are affected by poor data, such as deadlines not being met or quality expectations falling short, reputational damage can occur. If services or products are based on inaccurate data, they will not meet user needs, leading to a loss of trust and customers.

6. Privacy/Security Issues

In cybersecurity scenarios, improperly validated data or poorly managed access controls can lead to leaks or unauthorized access, exposing sensitive information.

These are examples of how poor data validation, and thus poor data quality, can have cascading negative effects, from affecting downstream users to reputational damage and customer loss. For example, if poor data leads to bad product quality or delayed production, this can damage the company's reputation and lead to lost customers.

One notable example occurred in 2018 when Samsung Securities (a stock trading arm of Samsung) experienced a \$105 billion USD "fat finger" data entry error. An employee accidentally entered "shares" instead of "won" (South Korean currency), leading to the distribution of 2.8 billion shares instead of 2.8 billion won.

Although this error was due to human data entry, similar issues can arise when working with sensors and automated data collection, where human involvement is minimal. However, even automatically collected data needs validation to ensure trustworthiness. Validating large volumes of data from sensors manually is infeasible, so it is essential to automate data validation within the system where data is collected.

Once data is properly validated, it becomes a powerful asset for driving informed decision-making. However, even the most accurate data is only valuable when it can be effectively interpreted. This is where data visualization plays a pivotal role, transforming validated data into actionable insights that help companies optimize their operations and strategies.

What is Data Visualization, and Why is it Important

Definition of Data Visualization

Data visualization is the graphical representation of information and data, using visual elements such as charts, graphs, maps, and infographics. By translating complex datasets into a visual format, data visualization enables a clearer understanding of patterns, relationships, and trends within the data. This process leverages visual cues like colors, shapes, and sizes to make data more accessible and comprehensible, facilitating quicker and more effective analysis.

Importance of Data Visualization in Decision-Making

In the context of decision-making, data visualization is a critical tool that transforms raw data into actionable insights. For production companies, where decisions often need to be made swiftly and accurately, visual representations of data are invaluable. Instead of sifting through lengthy reports or large datasets, decision-makers can quickly grasp key information through well-designed visualizations. This leads to more informed and timely decisions, minimizing the risks of misinterpretation or oversight that can occur when dealing with raw data.

Benefits of Visualizing Data: Trends, Patterns, and Insights

For production companies, the ability to visualize data offers several significant benefits:

1. Identifying Trends

Data visualization makes it easier to identify trends over time. For instance, a production company can track the efficiency of a manufacturing line or the fluctuation in demand for a product across different seasons. Recognizing these trends allows for better forecasting and more strategic planning, ultimately leading to cost savings and increased profitability.

2. Spotting Patterns

Visualizations help uncover patterns that might not be immediately apparent in raw data. For example, a heat map could reveal which machinery tends to require more frequent maintenance, or a line graph might show a correlation between production output and staffing levels. Understanding these patterns helps production companies optimize their operations, reduce downtime, and improve overall efficiency.

3. Gaining Insights

Perhaps the most crucial benefit of data visualization is the ability to gain deep insights into various aspects of production processes. Whether it's understanding customer behavior, optimizing supply chains, or enhancing product quality, visualizations enable companies to dig deeper into their data and discover insights that can drive innovation and competitive advantage.

While data visualization transforms raw data into actionable insights, it relies heavily on the underlying infrastructure that processes, organizes, and delivers data in real time. This is where tools like Node-RED come into play. Node-RED acts as the bridge between data sources and visualization platforms, handling the complex tasks of data collection, processing, and automation before the data is presented in tools like Power BI and Grafana. The combination of real-time data processing and advanced visualization ensures that companies can make informed decisions based on the most up-to-date and accurate information available

Data Processing: Node-RED

Introduction and Features of Node-RED

Node-RED is an open-source, flow-based development tool originally designed by IBM to wire together hardware devices, APIs, and online services in innovative ways. It is particularly useful for production companies looking to streamline data acquisition and automation processes. Node-RED provides a visual editor that allows users to wire together various nodes—representing different data sources, processing functions, and outputs—into workflows capable of handling tasks ranging from sensor data collection to complex automation.

Key features of Node-RED for production companies include:

1. Visual Programming Interface

Node-RED's drag-and-drop interface simplifies the creation of workflows, making it accessible for engineers and technicians who may not have extensive programming experience.

2. Integration with IoT Devices

Node-RED excels in integrating with industrial IoT (IIoT) devices, enabling real-time data collection and processing from production equipment, sensors, and machinery.

3. Automation Capabilities

Production companies can automate tasks such as monitoring machine health, adjusting operational parameters, and triggering alerts when specific conditions are met.

4. Extensive Library of Nodes

Node-RED's library includes a wide variety of pre-built nodes that support different protocols and services, making it easy to connect with existing production systems and external APIs.

5. Open Source and Flexible

As an open-source tool, Node-RED offers a cost-effective solution that can be customized to fit the specific needs of a production environment.

By integrating Node-RED's data processing capabilities with visualization tools like Power BI and Grafana, companies can ensure that real-time data is processed efficiently and presented in a way that enables swift, data-driven decisions. This seamless connection between data processing and visualization guarantees that decision-makers always have the most up-to-date information at their fingertips.

Data Visualisation: PowerBI & Grafana

In a production environment, different tools serve different visualization needs. Power BI is ideal for creating business-centric reports and dashboards, while Grafana excels at monitoring real-time data from IoT devices. Together, these tools provide comprehensive data visualization options for both strategic decision-making and operational monitoring.

Introduction and Features of Power BI

Power BI, developed by Microsoft, is a robust business analytics tool designed to deliver interactive visualizations and actionable insights. For production companies, Power BI offers the ability to connect to a wide range of data sources – ranging from manufacturing databases to supply chain systems – allowing users to transform raw data into meaningful reports and dashboards. Its user-friendly interface and seamless integration with other Microsoft products, such as Excel and Azure, make it a valuable asset for production managers and decision-makers who need to quickly analyse and respond to data-driven insights.

Key features of Power BI for production companies include:

1. Drag-and-Drop Interface

Production managers and team members can easily create reports and dashboards without needing deep technical expertise, streamlining the data analysis process.

2. Customizable Visuals

Tailor visualizations to monitor specific production metrics, such as machine efficiency, output rates, or quality control measures.

3. Data Modeling and Transformation

Power BI enables complex data modeling, allowing companies to track and analyze relationships across various production stages, from raw materials to finished goods.

4. Real-Time Analytics

Power BI provides real-time data updates, ensuring production teams have the latest information on manufacturing performance, helping to quickly address any issues or inefficiencies.

5. Integration with Microsoft Ecosystem

Seamless integration with tools like Excel and SharePoint enhances data connectivity and collaboration across different production departments.

Introduction and Features of Grafana

Grafana is an open-source platform known for its strength in visualizing time-series data, making it particularly useful for monitoring and optimizing production processes. Production companies can leverage Grafana to visualize data from various sources, providing a comprehensive view of operational metrics such as equipment performance, production line efficiency, and downtime tracking. Grafana's flexibility allows for creating customized dashboards tailored to the specific needs of a production environment.

Key features of Grafana for production companies include:

1. Data Source Agnostic

Grafana supports a wide range of data sources, from industrial IoT devices to databases like InfluxDB and Prometheus, enabling comprehensive monitoring across the entire production process.

2. Custom Dashboards

Create dashboards that visualize critical production metrics, helping teams monitor performance and make data-driven decisions to improve efficiency.

3. Alerting System

Set up alerts for critical production parameters, such as equipment failures or deviations from standard operating conditions, ensuring timely responses to potential issues.

4. Plugins and Extensions

Grafana's extensive plugin ecosystem allows production companies to expand functionality with additional visualization tools and integrations, enhancing their data analysis capabilities.

5. Collaboration Features

Shared dashboards and annotations enable teams to collaborate more effectively, ensuring everyone has access to the same real-time data and insights.

While tools like Power BI and Grafana help production companies visualize and analyze data in real time, more advanced applications can leverage this data for automated decision-making. Closed-loop control systems represent the next stage in this process, where data is not only monitored but also used to automatically adjust and control production systems in real time. By integrating data processing, visualization, and control systems, production environments can achieve higher efficiency, reduced downtime, and more precise operational control. The following section outlines the architecture of such a closed-loop control application, highlighting how data flows through various stages to enable autonomous system adjustments.

High level architecture of a data-driven closed loop control application

In modern data intensive applications, data is everything and everything is data. In the fast-moving world of data intensive IoT applications, the importance of data quality cannot be undermined. Without achieving quality of data at a certain level required by a use case the Return on Investment (ROI) in IoT applications is almost none. Thus, it is important that while developing any data intensive application, the quality of data should be given the importance it deserves. In IoT applications data produced, processed and used for any decision-making passes through several intermediate steps. These intermediate steps can be carried out at edge devices, on-prem, or in the cloud.

Furthermore, the growing competition by cloud service providers to make on demand infrastructure affordable and accessible have opened many doors of opportunities for small and medium enterprises, ambitious to digitalize their operations. One such area where most of the companies are investing is IoT based production optimization. There are several methods of production optimization, an important method is to introduce IoT based closed-loop control systems to reduce any manual involvement in production operations.

There are numerous closed loop applications that we come across regularly, a classic example of closed-loop control is temperature control system installed in most homes, a temperature sensor on periodic intervals calculates the current room temperature and then sends it to a regulatory system. The regulatory system on the basis of current temperature readings sets the new temperature by switching on/off the heater. Cruise control system that maintains the speed of vehicles and HVAC systems that maintain the temperature in the vehicles are also some very common applications. In industrial applications closed-loop control systems are used to control physical properties and quantities, such as pressure (as shown in Fig-1), temperature, humidity, oxygen level, pH level, density, boiling point, color, volume, melting point etc.

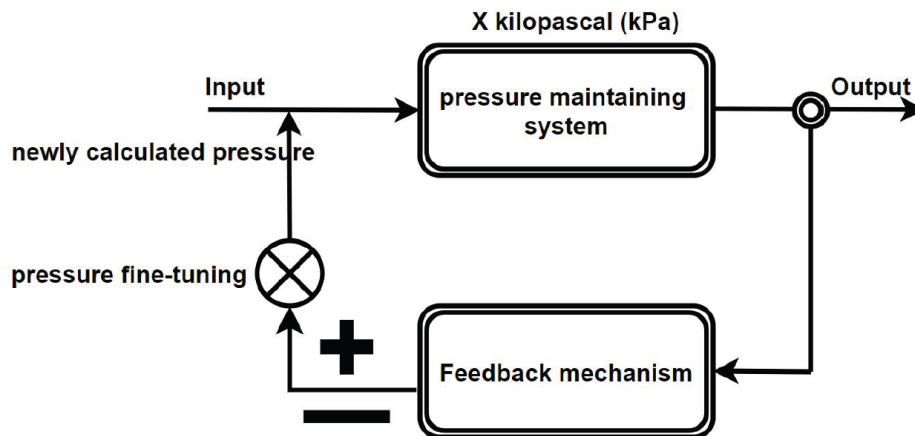


Fig. 1: A closed-loop system maintaining the pressure.

Closed-loop control systems in industries usually have two major components, i) the software part that is comparing the current system state with the required system state and sending the newly calculated state to some hardware (electro-mechanical) components actuators etc. ii) The electro-mechanical component that upon receiving the recalibration commands from the software mechanically adjusts the hardware to keep the system in the required state. Often closed-loop applications are supplemented with machine learning models and data analytics carried out in the cloud. This needs data to be moved to the cloud, the volume, velocity and variety of data moved to the cloud is very application specific.

Moreover, in situations where companies have limited bandwidth available due to higher regional tariffs or high network traffic it is very important to select data transmission methods suitable for specific application needs. This use case focuses on the software part of the closed-loop control system and proposes an application architecture that is not only suitable for low bandwidth environments, but also tailored to fit several IoT and closed-loop control use cases.

Why do companies invest in closed-loop applications: Companies looking to digitalize their production lines and day to day operations opt for cloud platform. One of the focuses of digitalization is Closed-loop control system. Closed-loop control systems not only increase productivity but also reduce possible down-time and can be very effective when coupled with predictive maintenance systems.

Closed-loop applications implementation challenges: Closed-loop control systems are data dependent, this data can be obtained through various protocols, APIs etc. However, selection of specific methods, protocols, APIs is very application specific. There are various pros and cons associated with each technique, therefore, providing details of each of them are beyond the scope of text here. A closed-loop control system processes data required to adjust and control the system state. To continuously keep track of the state and adjust the input(s) accordingly, a closed-loop control system needs to process streaming data. Due to various security, compliance, geographical, business and company related internal and external regulations directly accessing on-prem infrastructure and network is often

restricted. Therefore, in such cases it is important to select suitable technologies to overcome such restrictions.

The following are some options which can be considered for exchanging data between cloud infrastructure and on-prem set-up.

Representational State Transfer (REST)

- Request/Response architecture (polling can help).
- If direct requests to on-prem networks are not allowed, then a middleware, API gateway or API proxy can be used.
- REST is suitable for create, retrieve, delete, or update operations on resources and not for continuous stream required by many IoT applications.

WebSocket

- Allows full-duplex communication over a single TCP connection.
- Ideal for low-latency, high-frequency communication.
- Suitable for point to point (one to one) communication.

Message Queuing Telemetry Transport (MQTT)

- Lightweight, suitable for resource constrained environments and supported by millions of cheap devices
- Multiple QoS modes to ensure message delivery for different application requirements
- No direct communication between clients for data exchange, useful where direct communication is not possible due to various internal or external regulatory constraints.

From the aforementioned data transfer and data exchange technologies MQTT is better suited for closed loop control applications where direct connectivity is not possible due to strict security and compliance reasons to any on prem infrastructure, and the size of data and frequency of events is not hard real time. An MQTT broker in the middle in this approach by default hides the source and destination data sources without needing any extra efforts. REST is not suitable for streaming data and Web sockets are more suited for point-to-point communication.

Therefore, due to the suitable nature for resource constrained environments and due to one-to-many capabilities MQTT should be the preferred approach to transmit and receive on-prem event data to control applications and hardware.

High level Architecture

The sample high level architecture consists of two major parts. One on-prem control systems, this part directly interacts with the closed loop system hardware to control and stabilize the physical process. On-prem control system receives commands from an

external system that we call off-prem control system. Due to various regulatory, security and compliance requirements, on-prem systems are always not directly exposed to any off-prem applications and systems refer to the Fig.-2. Commands and instructions received by the on-prem system are converted to electrical signals, which are then used to control different types of sensors, machinery or processes.

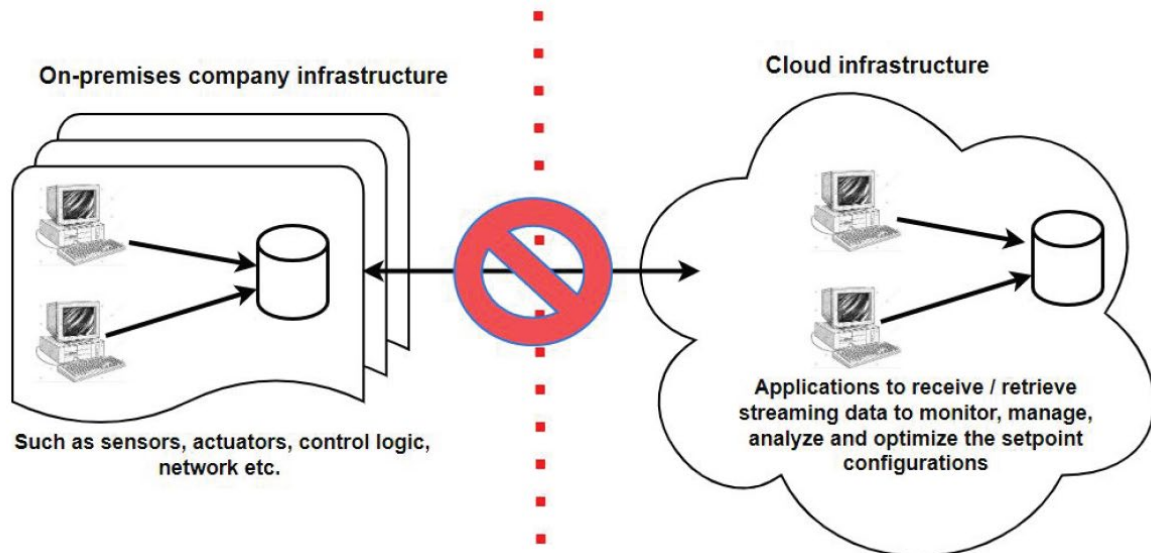


Fig. 2: Restrictions on direct data access.

Off-prem or cloud part of the closed loop control system receives current status information of the process and after analyzing the current status data, triggers some rules and sends back newly calculated information to the on-prem system. Off-prem components are typically hosted in the cloud, however, there are no restrictions whatsoever as individual application components can be hosted anywhere and, on any hardware, meeting the minimum individual component hosting requirements.

For a real-world industrial example of a closed loop application, assume a consumer appliances assembly plant where washing machines, air-conditioners, fans, ovens and vacuum cleaners are being painted. Different appliance components throughout the assembly line receive different chemical and color treatments. These chemicals and paints are applied based on the type of component, model of the component, structure of components (plastic or metal), protection required (for ex. from rust) etc.

Assuming that components moving on assembly line can arrive in different order, then dynamic chemical and paint selection requires that a system reading the part number can quickly adjust the temperature, pressure, volume and humidity (if required) of the chemical or paint applied to that component. To keep the assembly line moving, this dynamical tuning of temperature, pressure, volume, humidity or any other physical property at any stage requires two things to be known. The first thing existing status of the property such as current pressure level, and second thing is the required pressure level. The current status of a physical process and required status of a physical process lets the closed loop application calculate the difference and then send commands to the hardware

(electromechanical devices) responsible to physically adjust the process to maintain it in a stable state. The current values of physical properties such as pressure, temperature, volume etc. are continuously sent as an event stream. Each event is sent to the off-prem application hosted in the cloud. Where updated properties are calculated and sent back to the on-prem system for further action.

Application Components

The following application components can be built to implement a solution for an assembly plant mentioned above.

1. Internal Application components

i. Internal infrastructure

Any hardware or software that directly interacts and controls the electromechanical components used to stabilize the processes is considered as part of internal infrastructure. Thus, their working principles and details are not mentioned here. It is up to the organization how it implements and manages its internal infrastructure.

ii. Internal infrastructure façade

The façade component can be used to shield the organization’s internal infrastructure and prevents the external organizational components to directly communicate with the internal hardware and to regulate the closed loop system. It also provides a transparent interface to the external applications to receive and send data without knowing about any application logic. Furthermore, this façade also serves as a gateway to transform the internal data received into JSON string and external data back to the format expected by organization’s internal infrastructure and application.

2. External Application components

External applications components are divided into three major parts as follows.

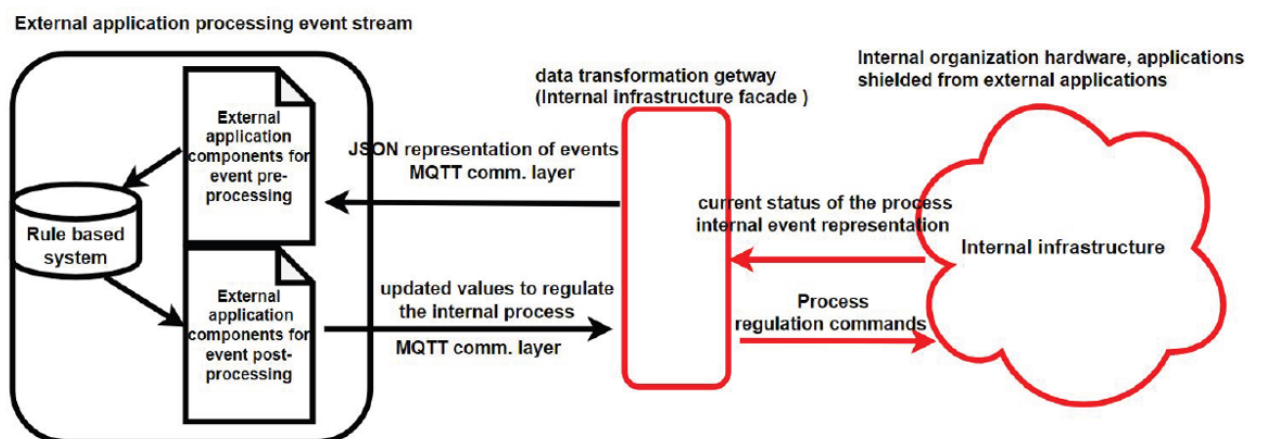


Fig. 3: Distribution of processing between on-prem and off-prem infrastructure.

i. Event Pre-processing Component

This component as shown in the figure is based on an MQTT client and custom logic. The client receives an MQTT message as a string. This string is a representation of an event in JSON encoded format. This JSON encoded string is then supplied to a custom converter that converts each MQTT message to an instance of an Event class. If it is required each event instance can be stored in the database for any further analysis, future use case or machine learning purpose. The current state of the closed loop application process i.e. temperature and humidity values are extracted from each event instance and supplied to another application component that keeps the current system state.

ii. Rule based System

The Rule based system is the core component of the external application. Rule based system should contain all the rules required to keep the closed loop system and the process being controlled in a stable state. The number of rules and the complexity of each rule varies on the process being stabilized and parameters being monitored. In the absence of a rule-based system, simple rules can be created using if-else or switch statements by features common in all modern general purpose programming languages. However, switch statement is not great to handle complex rules, and implementing complex rules using if-else statement could lead to potential logical errors difficult to spot and debug and may lead to classic if-else hell issues. Thus, when the rules are complex and the list of rules is expected to grow then a rule-based system, and a dedicated library to handle that is a better option.

iii. Event Post-processing Component

The event post-processing component receives the information when a particular rule is triggered by the rule-based system. It then uses the transformation logic to translate the rule to an MQTT message expected by the Internal infrastructure gateway. An MQTT client is then used to send this message to the Internal infrastructure gateway/façade.

The aforementioned high-level architecture covers the basic building blocks of a closed loop control application for a particular use case. The number of components, protocols used, and control logic can be distributed between components to suite any other use case. Moreover, timeliness, latency, frequency of events, size of data exchange required for the control between components, security, compliance can result in a different architecture as well.

Conclusions & recommendations

Data validation is a critical process that ensures accuracy, reliability, and fitness-for-purpose of data, which is fundamental for making informed decisions in production environments. Without proper validation, data quality diminishes, leading to incorrect conclusions, increased operational costs, and potential reputational damage. In production settings, where precision and efficiency are paramount, decisions based on faulty data can result in inefficiencies, equipment failures, and lost revenue. By implementing automated data validation processes, companies can ensure that their data remains trustworthy and supports reliable decision-making.

Once data is validated, data visualization becomes a powerful tool for turning complex data into actionable insights. Visualization aids in identifying trends, discovering patterns, and extracting valuable information, all of which support better decision-making, improved efficiency, reduced costs, and increased profitability.

For workflow automation and real-time data handling, **Node-RED** is particularly well-suited for production companies needing to integrate with IoT devices and automate processes. In contrast, **Jupyter** is ideal for environments that require advanced data analysis, documentation, and collaboration, especially when combined with machine learning for predictive analytics. The choice between Node-RED and Jupyter depends on the specific goals of the company – whether they are focused on real-time automation or deep data analysis.

When it comes to business intelligence and reporting, **Power BI** stands out for its seamless integration with Microsoft products, making it ideal for companies focused on business analytics. On the other hand, **Grafana** is better suited for real-time monitoring of equipment performance and cost-effective visualization. The decision between Power BI and Grafana should align with the company’s needs, whether they prioritize in-depth business analytics or operational monitoring.

Furthermore, when data is put into application use, it is important to consider what data transfer, data processing, and data storage technologies are relevant. In the case of closed loop control application, a use case is provided that suggests some tools and data transfer techniques that can come handy in similar scenarios.

Kasper Stens Honoré
M.Sc.
Technical Project Manager
FORCE Technology
ksho@forcetechnology.com
+45 4262 7042

Mads Johansen
M.Sc.
Specialist
FORCE Technology
majh@forcetechnology.com
+45 4262 7531

Ahmed Khan Leghari
PhD.
Senior Specialist
FORCE Technology
ahkl@forcetechnology.com
+45 4262 7160

November 2024