P. Kannisto & D. Hästbacka, "Data Autonomy in Message Brokers in Edge and Cloud for Mobile Machinery: Requirements and Technology Survey",
in 27th IEEE International Conference on Emerging Technologies and Factory Automation, 2022.
doi:10.1109/ETFA52439.2022.9921693

Data Autonomy in Message Brokers in Edge and Cloud for Mobile Machinery: Requirements and Technology Survey

Petri Kannisto
Tampere University
Tampere, Finland
ORCID: 0000-0002-0613-8639

David Hästbacka
Tampere University
Tampere, Finland
ORCID: 0000-0001-8442-1248

***Abstract*—The future data-driven manufacturing ecosystems build upon data spaces where each participant controls how its own data are utilized. This goal is equally important in machinery where the networks comprise machine fleets and the data use cases range from local edge to various cloud systems and digital business integrations. These systems of systems require efficient, scalable data streaming with decoupled (e.g., publish-subscribe) platforms that enable the flexible connection of data producers and consumers in heterogeneous networks. This paper describes a work in progress about data autonomy, sovereignty, and trust in message brokers in machinery, aiming to contribute to initiatives, such as Gaia-X and International Data Spaces (IDS). First, the paper identifies requirements for platforms and communication that span edge and cloud. Second, it presents a technology survey about data autonomy in open, Internet-cabable brokers. These include Advanced Message Queueing Protocol (AMQP), MQ Telemetry Transport (MQTT), Apache Kafka and Apache Pulsar. It appears that there is little research about data autonomy in brokers and MQTT has the strongest base. The work continues to develop data autonomy into a message broker.**

***Index Terms*—Message-oriented Middleware, Data-driven Systems, Industry 4.0**

I. Introduction

The business value enabled by data-driven methods spans mobile machinery and machine fleets similar to manufacturing. Being data-driven refers to acting on data rather than opinions or intuition [1]. Data enable the generation of information and knowledge in machine fleets [2]. Due to huge data amounts, the fleets should take advantage from cloud computing for scalability and easy integration to other services and stakeholders [3]. Still, the cloud should be complemented with edge computing to bring the utilities closer to those clients that generate and use the data [4]. The fleets, cloud, and edge form complex, multiparty systems of systems (SoS [5]).

The multiparty nature of the data-driven business causes challenges regarding data autonomy, sovereignty, and trust. The parties have various roles (e.g., machine manufacturer, operator, or maintenance provider) and operate systems and software from various vendors. In such an environment, the control of data usage is currently challenging. Data autonomy means letting the data owner control data usage, whereas sovereignty refers to applying the laws of data origin. Furthermore, the parties should be able to trust each other. To meet these requirements, Gaia-X has arisen to provide an infrastructure with trust and identity, data catalogues, and sovereignty, guaranteeing compliance with common rules [6]. For concrete software solutions, Industrial Data Spaces Association (IDSA) has specified a reference architecture to consider security, certification, and governance from business requirements to software implementations [7]. However, Gaia-X and IDSA in machinery remain unexplored.

This paper describes work in progress on the management of data autonomy in machinery ecosystems. First, the requirements of communication are resolved in a workshop of practitioners and researchers. Second, a technology survey is provided on the data space readiness of message broker technologies. These include the open standards Advanced Message Queueing Protocol (AMQP) and MQ Telemetry Transport (MQTT) as well as the open-source products Apache Kafka and Apache Pulsar. Thus, the research objectives are:

- RO1: Identify requirements for communication in machine fleets where computation spans both edge and cloud
- RO2: Survey the capabilities of open message broker technologies to manage data autonomy

The long-term goal, beyond this paper, is to combine the advantages of a message broker with data autonomy. This paper provides requirements and a survey for the foundation of the solution. While this study focuses on machinery, it contributes even to factory automation, which has similar requirements related to data utilization and repeatedly changing environments [8].

Next, Section II explains the challenges of data exchange and utilization in edge and cloud. Section III elaborates requirements gathered from practitioners and researchers (RO1). Section IV selects the brokers and compares these, followed by a literature search for data autonomy in Section V (RO2). Finally, Section VI concludes the paper.

This paper was financially supported by the project ”Future Electrified Mobile Machines” (FEMMa) funded by Business Finland (diary 2801/31/2021). The authors express their gratitude to the funder, partners, and supporters.

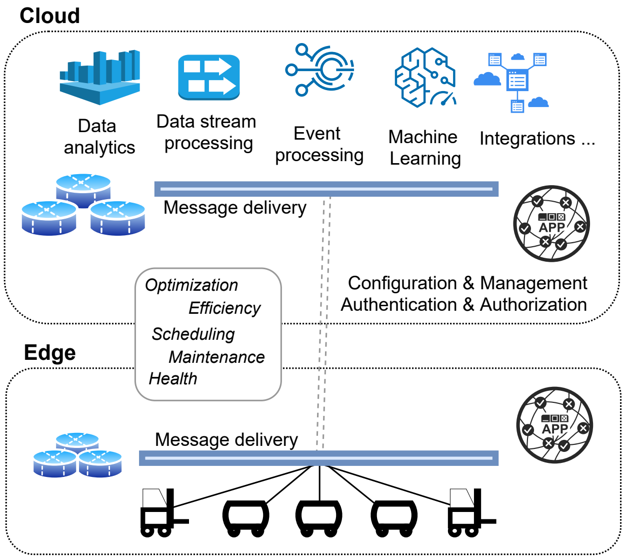

Fig. 1. The messaging infrastructure is paramount in delivering data for edge and cloud computing.

II. Messaging for Edge and Cloud

Fig. 1 illustrates how multiple services provide added value in the cloud and edge in a data-driven machinery ecosystem. A data platform as a shared communication medium is desirable, as it decouples individual entities from each other and improves the flexibility of integrations. On the other hand, it can reduce performance, introduce a single point of failure and shift the control of data and integrations to a layer outside of the systems. Typically, it is the cloud that provides services, such as analytics, processing, and machine learning. The edge provides local services for the machine fleet, possibly processing some data locally or filtering the data transferred to the cloud [4].

A message broker (the bar in both edge and cloud) can connect the data sources and services involved. The broker can solve many of the communication-related issues. It can provide data routing over Internet, removing geographical boundaries. It can be resilient, i.e., enable the network to remain operational even if a data provider or consumer is occasionally unreachable. It can be asynchronous so that each node can produce or receive data in its own pace [9]. Finally, it can scale as the amount of traffic and number of network nodes grow. This means that from the communication viewpoint, a message broker provides a solid foundation.

III. Identification of Data Exchange Requirements

A. Questions and Replies

To resolve concrete requirements for data utilization, an online workshop was organized with representatives from four machine manufacturers, a research institution and a university, each from Finland. They were asked questions, after which they replied with electronic post-it notes on a collaborative web platform. Then, the replies were discussed. The process delivered qualitative results regarding the requirements. The questions and respective replies are aggregated and elaborated in the following paragraphs.

a) Question 1: Types of Information: What kind of machine information is needed by other systems and operations in close to real time and in the long run?

- Measurements; efficiency, performance
- Task control, management, and planning
- Location; environment/surroundings information
- Health; maintenance history
- Battery state (if electric machine)

b) Question 2: Opportunities: What kind of external parties would you foresee making use of that data in joint operations? What would be the benefits?

- Information sharing between parties
- System-of-systems scope: efficiency, co-operative optimization, overall planning
- Scheduling between operations and service

c) Question 3: Challenges: What are the current challenges in sharing and managing data? What kind of security requirements arise?

- Heterogeneity of and distribution to multiple data sources
- Appropriate metadata
- Data quality
- Information security
- Trust, data ownership, storing sensitive data in an external cloud, and sovereignty
- Common data sharing practices; distinct business models
- Data jealousy
- Identification of what data should and could be shared
- Risk analysis on company and ecosystem level
- Responsibilities in decision-making based on complex data analysis, including external data sources and services

B. Outcome from Replies

The workshop provided insight about the needs and visions regarding machinery ecosystems. The participants see potential in generating benefit from data and their sharing among business partners. Still, there are clear challenges regarding both business aspects and technology.

The results contain a few challenges that could and should be solved with an appropriate message broker. From the communication viewpoint, question 1 resulted in requirements to the information models of messaging. These are technological but not related to the broker but rather to message structures. Question 2 resulted in envisioned non-technical benefits. Among the replies to question 3, the following challenges can be met with the broker at least partially:

- *Heterogeneity of data sources* can be partially mitigated, but this requires common information models too [10]
- *Distribution to multiple data sources* is no problem technically if the medium can route messages over Internet
- *Information security* is a process but also requires, e.g., data encryption, usage control, and integrity control

- *Trust*, *data ownership*, and *sovereignty* can be supported with the methods of managing data autonomy

Of these challenges, trust, data ownership, and sovereignty are within the scope of this article.

IV. Internet-capable Message Brokers

A. Selection of Messaging Technologies

To suit for edge and cloud computing, the messaging infrastructure must support the following features:

- Either open standard or open-source product
- Routing over Internet
- Publish-subscribe communication
- No restrictions to data serialization

These features are fulfilled by a number of broker technologies, including AMQP [11], [12], MQTT [13], Apache Kafka [14], and Apache Pulsar [15]. AMQP and MQTT are open standards that specify a broker without any data storage. In contrast, Kafka and Pulsar are open-source products that can even store data for a certain period. The two most recent versions of AMQP, 0-9-1 and 1.0, are competing and incompatible. 1.0 only specifies a data link, but there are broker implementations too. Despite differences, each technology enables publish-subscribe and can enhance scalability in data streaming compared to request-response technologies.

Multiple open technologies provide publish-subscribe communication but remain excluded from this work. Data Distribution Service (DDS) provides a brokerless publish-subscribe-capable channel. DDS excels in limited subnets but does not scale to Internet due to multicasting. ZeroMQ is another brokerless technology with the respective limitations.

B. Current Capabilities of Brokers

As trust and sovereignty build upon security, Table I shows how the selected broker technologies currently support security features. The scope of encryption can be limited to the communication channel (with Transport Layer Security, TLS) or cover the entire path from publisher to subscriber. Authentication can use multiple means, some of the common including username and password, JSON Web Token (JWT), and OAuth 2. Each can be either supported directly or via the method layer Simple Authentication and Security Layer (SASL) that enables a range of techniques. For access control, a publish-subscribe system gains benefit if the resolution can reach topics.

It appears that the brokers provide largely similar features in terms of security. Each provides a range of authentication methods and at least channel encryption, whereas Pulsar can even encrypt end-to-end. Topic-level access control is lacking from each standard (i.e., AMQP and MQTT) although MQTT 5 suggests this non-normatively and at least HiveMQ has an implementation. Both Kafka and Pulsar support this. While the reviewed features are essential, they only form a foundation for data autonomy. There must be additional techniques to control data usage in a multi-party network.

TABLE I
Security features supported by message broker technologies

Feature	AMQP	A. Kafka	A. Pulsar	MQTT
–Encryption				
Channel (TLS)	✓(a)	✓	✓	✓
End-to-end	-	-	✓	-
–Authentication				
SASL (subset or all)	✓	✓	✓	✓
Username & password	✓(b)	✓	✓	✓
JWT	✓(b)	✓	✓	✓(b)
OAuth 2	✓(b)	✓	✓	✓(b)
–Access control				
Topic level	-	✓	✓	- (c)

(a) AMQP 1.0 only; 0-9-1 has implementation, at least RabbitMQ
(b) Via SASL
(c) MQTT 5 suggests this non-normatively; at least HiveMQ implements

V. Literature About Managing Data Autonomy

A. Search Method

To study the security potential of the brokers, a literature search was performed. The search was conducted with the citation index Scopus, searching in article title, keywords, and abstract. Scopus was considered wide and therefore to provide reliable results. The following searches were included.

a) Usage control: This covers access control and its superset usage control or UCON [16]. The search string is ’”usage control” OR ”access control” OR authorisation OR authorization’.

b) Goals of managing data autonomy: The goals are data autonomy, data sovereignty, and trust, thus the search string is ’”data autonomy” OR ”data sovereignty” OR trust’.

c) Platforms: The leading platforms for managing data autonomy are Gaia-X and international data spaces. The search string is ’”data space” OR ”gaia x”’.

For the final search string, the aspects explained above were combined with each technology. For the technologies, the search strings are ’amqp’, ’mqtt’, ’apache AND kafka’, and ’apache AND pulsar’. This would finally be combined with the condition ’TITLE-ABS-KEY’ to search in the title, abstract, and keywords. For example, to search for access control regarding Apache Pulsar, the string is:

```
TITLE-ABS-KEY (apache AND pulsar AND ("data
space" OR "gaia x"))
```

The search results were taken into a spreadsheet and inspected to decide whether to include each article. The searches returned only the following document types, each considered relevant: conference paper, article, and conference review.

B. Search Results

Table II shows the number of results along with excluded articles if any. The exclusions occurred to items that only introduce a conference event. The search time was June 2022.

It appears that only MQTT has received significant research regarding the themes relevant to data autonomy. This is understandable for Kafka and Pulsar as they lack a user base and tradition similar to MQTT. Still, AMQP can be considered historically popular but still has almost no research

TABLE II
RELATED PUBLICATIONS (EXCLUDED ITEMS IN PARENTHESES)

Theme	**AMQP**	**A. Kafka**	**A. Pulsar**	**MQTT**
Usage control	5	4	0	73 (6)
Data autonomy	0	1	0	20 (4)
Platforms	0	0	0	1

for the themes. Presumably, the difference between AMQP and MQTT stems from their conventional use case. MQTT was originally an IoT protocol, whereas AMQP was developed for large-scale enterprise systems in the finance industry [17]. IoT systems may require more of access control as the environment is more heterogeneous and the nodes enter or leave the network more often. Overall, the research base of MQTT provides by far the best material for further developments. Still, even MQTT has almost no research regarding autonomy-related platforms, which is a research gap. On the other hand, grey literature (blogs, etc.) could be searched for more results.

For an overview, there have been multiple studies to extend access control in MQTT. Calabretta et al. [18] have studied the application of tokens for authentication and authorization. La Marra et al. [19] propose an approach to include usage control. Nast et al. [20] propose an International Data Spaces (IDS) adapter. Colombo et al. [21] have studied the application of Attribute-based Access Control (ABAC). Nichols [22] describe how trust schemata can define access rules.

VI. CONCLUSIONS AND FUTURE WORK

In conclusion, this article described work-in-progress research with two objectives: RO1 about the requirements of data-driven edge and cloud computing in mobile machinery, and RO2 about the capabilities of message broker technologies for data autonomy. While mobile machinery has its special characteristics, particularly mobility, the problems are generalizable within manufacturing, which is data-intensive and typically requires repeated re-configurations in the facilities.

For RO1, some of the identified requirements are business related or organizational and only a few can be completely solved with technology. The results are qualitative due to a low number of participants and restricted to Finland.

For RO2, the results were mixed. The brokers appeared mostly similar regarding their functionality for publish-subscribe communication. However, only MQTT has received significant research regarding extensions for data autonomy and related topics. Still, it can be argued that respective solutions would work with other publish-subscribe-capable technologies too, and MQTT is an old technology compared to Apache Kafka and Pulsar, which can still deliver more of significant research. Overall, the results revealed only one message-broker-related article for data space connections [20]. Therefore, the area should be studied more to enable the brokers to operate with initiatives such as IDS and Gaia-X. Still, grey literature (blogs, etc.) could provide more results.

Based on this study, the research will continue on an actual data autonomy solution for mobile machinery based on a message broker. This paper provided both requirements and a survey of the technologies to build upon. Future studies will include architectural considerations for cloud and edge as well as proofs of concept based on practical use cases. The solution should support the goals of Gaia-X [6] and IDSA [7].

REFERENCES

[1] C. Anderson, *Creating a Data-Driven Organization*. O'Reilly, 2015.

[2] P. Kannisto, D. Hästbacka, and S. Kuikka, "System architecture for mastering machine parameter optimisation," *Comput. Ind.*, vol. 85, pp. 39–47, 2017, DOI: 10.1016/j.compind.2016.12.006.

[3] P. Kannisto and D. Hästbacka, "Cloud-based management of machine learning generated knowledge for fleet data refinement," in *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, ser. Commun. Comput. Inf. Sci., A. Fred *et al.*, Eds. Springer, 2019, pp. 267–286, DOI: 10.1007/978-3-319-99701-8_13.

[4] W. Z. Khan, E. Ahmed, S. Hakak, I. Yaqoob, and A. Ahmed, "Edge computing: A survey," *Future Gener. Comput. Syst.*, vol. 97, pp. 219–235, 2019.

[5] C. Keating, R. Rogers, R. Unal, D. Dryer, A. Sousa-Poza, R. Safford, W. Peterson, and G. Rabadi, "System of systems engineering," *Eng. Manag. J.*, vol. 15, no. 3, pp. 36–45, 2003.

[6] "GAIA-X: Driver of digital innovation in Europe," 2020, URL https://www.data-infrastructure.eu/GAIAX/Redaktion/EN/Publications/gaia-x-driver-of-digital-innovation-in-europe.pdf [Visited 12 Jul 2022].

[7] "IDSA reference architecture model," 2019, URL https://internationaldataspaces.org/wp-content/uploads/IDS-Reference-Architecture-Model-3.0-2019.pdf [Visited 12 Jul 2022].

[8] S. Dumss, M. Weber, C. Schwaiger, C. Sulz, P. Rosenberger, F. Bleicher, M. Grafinger, and M. Weigold, "EuProGigant -– a concept towards an industrial system architecture for data-driven production systems," *Procedia CIRP*, vol. 104, pp. 324–329, 2021.

[9] P. T. Eugster, P. A. Felber, R. Guerraoui, and A.-M. Kermarrec, "The many faces of publish/subscribe," *ACM Comput. Surv.*, vol. 35, no. 2, p. 114–131, Jun. 2003.

[10] P. Kannisto, D. Hästbacka, T. Gutiérrez, O. Suominen, M. Vilkko, and P. Craamer, "Plant-wide interoperability and decoupled, data-driven process control with message bus communication," *J. Ind. Inf. Integr.*, vol. 26, p. 100253, 2022, DOI: 10.1016/j.jii.2021.100253.

[11] "AMQP version 0-9-1," 2008, URL http://www.amqp.org/specification/0-9-1/amqp-org-download [Visited 17 May 2022].

[12] "OASIS AMQP 1.0," 2012, URL http://docs.oasis-open.org/amqp/core/v1.0/os/amqp-core-complete-v1.0-os.pdf [Visited 17 May 2022].

[13] "MQTT 5.0," 2019, URL https://docs.oasis-open.org/mqtt/mqtt/v5.0/os/mqtt-v5.0-os.html [Visited 17 May 2022].

[14] "Apache Kafka," URL https://kafka.apache.org/ [Visited 17 May 2022].

[15] "Apache Pulsar," URL https://pulsar.apache.org/ [Visited 17 May 2022].

[16] R. Sandhu and J. Park, "Usage control: A vision for next generation access control," in *Computer Network Security, MMM-ACNS 2003*. Springer Berlin Heidelberg, 2003, pp. 17–31.

[17] J. O'Hara, "Toward a commodity enterprise middleware," *Queue*, vol. 5, no. 4, p. 48–55, May 2007.

[18] M. Calabretta, R. Pecori, M. Vecchio, and L. Veltri, "MQTT-auth: a token-based solution to endow MQTT with authentication and authorization capabilities," *J. Commun. Softw. Syst.*, vol. 14, no. 4, pp. 320–331, 2018.

[19] A. La Marra, F. Martinelli, P. Mori, A. Rizos, and A. Saracino, "Improving MQTT by inclusion of usage control," in *Security, Privacy, and Anonymity in Computation, Communication, and Storage, SpaCCS 2017*. Springer International Publishing, 2017, pp. 545–560.

[20] M. Nast, B. Rother, F. Golatowski, D. Timmermann, J. Leveling, C. Olms, and C. Nissen, "Work-in-progress: Towards an International Data Spaces connector for the Internet of Things," in *2020 16th IEEE International Conference on Factory Communication Systems (WFCS)*, 2020, pp. 1–4.

[21] P. Colombo, E. Ferrari, and E. D. Tümer, "Regulating data sharing across MQTT environments," *J. Netw. Comput. Appl.*, vol. 174, p. 102907, 2021.

[22] K. Nichols, "Trust schemas and ICN: Key to secure home IoT," in *Proceedings of the 8th ACM Conference on Information-Centric Networking*, ser. ICN '21, 2021, p. 95–106.