Data Management over Cloud Edge Continuum in the context of AC^3

Dimitrios Amaxilatis
Spark Works Ltd.
Galway, Ireland
d.amaxilatis@sparkworks.net
[0000−0001−9938−6211]

Dimitrios Klonidis
Ubitech Ltd.
Athens, Greece
dklonidis@ubitech.eu
[0000−0002−4451−2062]

Adlen Ksentini
EURECOM
Sophia-Antipolis, France
adlen.ksentini@eurecom.fr

Vrettos Moulos
University of Piraeus
Piraeus, Greece
vrettos@unipi.gr

Souvik Sengupta
IONOS, SE
Karlsruhe, Germany
souvik.sengupta@ionos.com
[0000−0002−0056−1540]

Nikolaos Tsironis
Spark Works Ltd.
Galway, Ireland
ntsironis@sparkworks.net
[0009−0009−8084−613X]

Christos Verikoukis
University of Patras and ISI/ATHENA
Patras, Greece
cveri@isi.gr

***Abstract*—The proliferation of Internet of Things (IoT) technologies and the exponential growth of data have necessitated advanced data management solutions that can operate across distributed architectures. The AC^3 project introduces a Data Management Platform as a Service (DMPaaS), designed to address these challenges by providing a robust, scalable, and flexible platform bridging cloud and edge computing environments. This paper explores the core components of the AC^3 DMPaaS, emphasizing its dynamic data catalog, efficient data management processes, and the seamless integration of data connectors that facilitate data movement across the cloud-edge continuum. Through advanced metadata management, real-time data processing, and a hybrid cloud-edge data framework, the AC^3 DMPaaS supports complex data operations and analytics, ensuring data integrity and consistency. This comprehensive overview establishes the foundational elements of the AC^3 project's DMPaaS, showcasing its significant advantages for enterprises leveraging IoT data across diverse computing landscapes.**

***Index Terms*—Data Management, Cloud-Edge Continuum, Internet of Things, Data Sovereignty**

I. Introduction

In recent years, the proliferation of Internet of Things (IoT) technologies and the exponential data growth have necessitated more sophisticated data management solutions that operate across distributed architectures. The AC^3 project introduces a cutting-edge Data Management Platform as a Service (DMPaaS), designed to address these challenges by providing a robust, scalable, and flexible platform that bridges the gap between cloud and edge computing environments. This paper delves into the core components of the AC^3 project's DMPaaS, highlighting its innovative data catalog, efficient data management processes, and seamless integration of data connectors that facilitate fluid data movement across the cloud-edge continuum. The AC^3 DMPaaS is distinguished by its dynamic data catalog, which not only organizes a vast array of data types sourced from diverse IoT devices and applications but also enhances discoverability and accessibility. By employing advanced metadata management techniques, the data catalog ensures that data entities are easily searchable, thereby supporting complex data operations and analytics in real-time.

Moreover, the platform's data management capabilities are designed to handle the volume, velocity, and variety characteristic of modern data ecosystems. Through the integration of sophisticated data connectors, the AC^3 DMPaaS orchestrates the flow of data from edge devices to cloud-based systems and vice versa, optimizing data processing tasks for efficiency and reduced latency. These connectors are pivotal in maintaining data integrity and consistency across distributed environments, enabling real-time data processing and decision-making.

Finally, the paper explores how the AC^3 DMPaaS leverages a seamless cloud-edge data processing framework to execute data operations close to the data sources when necessary or in a centralized manner when appropriate. This hybrid approach allows for flexible deployment strategies that can dynamically adjust to varying workload demands, network conditions, and processing requirements, exemplifying a truly adaptive and resilient data management solution. This comprehensive overview establishes the foundational elements of the AC^3 project's DMPaaS, setting the stage for a detailed discussion on its implementation, operation, and the significant advantages it offers to enterprises seeking to harness the power of IoT data across diverse computing landscapes.

II. Related Work

The AC^3 project's development of an advanced DMPaaS underscores the strategic necessity of integrating a data catalog, optimized data management processes, and data connectors. This integration is pivotal to establish the AC^3 dataspace, which ensures secure and sovereign data sharing among diverse participants and stakeholders. Various architectural and governance components of the dataspace enable fluid data

This work is supported by the European Union's Horizon Program under the Agile and Cognitive Cloud edge Continuum management (AC3) project (Grant No. 101093129).

TABLE I
COMPARISON OF DATA MANAGEMENT SYSTEMS

Feature	Dataspace	Data Lakes	Data Lakehouse	Data Warehouses	Distributed Databases (DDB)	Federated Databases (FDB)	Multi-Databases (Multi-DB)
Primary Focus	Sovereign, interoperable data sharing	Storage of vast amounts of raw data	Combines Data Lakes and Data Warehouses	Single organization data analytics	Multi locations data distribution	Integration of multiple databases	Combination of data from different databases
Data Sovereignty	High	Low	Medium	Low	Medium	Medium	Medium
Interoperability	High	Low	Medium	Low	Medium	Medium	Medium
Trust and Security	High	Low	Medium	High	Medium	Medium	Medium
Data Governance	High	Low	Medium	High	Medium	Medium	Medium
Scalability	High	High	High	Medium	High	Medium	Medium
Data Integration	High	Low	Medium	Medium	Medium	Medium	Medium
Use Cases	Cross-organizational data sharing	Big data storage, analytics	Advanced analytics, ML, and BI	BI, reporting, analytics	Distributed data processing and storage	Cross-database queries and analytics	Cross-database queries and analytics
Examples	International Data Spaces, Gaia-X	Amazon S3, Azure Data Lake	Databricks Lakehouse Platform, Snowflake	Amazon Redshift, Google BigQuery	Google Spanner, Apache Cassandra	IBM InfoSphere Federation Server, Oracle GoldenGate	TIBCO Data Virtualization, Polybase

movement across the cloud-edge continuum, thereby embracing the concept of a dataspace and ensuring adherence to the EU Data Act[1]. This legislative framework aims to revolutionize data management, promoting accessibility, fairness, and interoperability, thus driving innovation and economic growth.

A dataspace is integral to a modern sovereign data management system, offering a secure and interoperable environment for data sharing among diverse stakeholders and participants. By ensuring data sovereignty, dataspaces empower data holders to control the terms and conditions of data reuse, thereby enhancing trust in data sharing mechanisms. The standardized protocols and frameworks embedded within dataspaces facilitate seamless data exchange, contract negotiation, and transfer management, akin to the role of the Internet Protocol in networking. This interoperability is crucial for the efficient deployment of advanced services and solutions across various sectors, fostering innovation and collaboration. Additionally, by adhering to international standards, dataspaces ensure legal certainty, data protection, and the integration of ethical guidelines, addressing the complexities of data governance in a globalized economy. Dataspaces offer several advantages over traditional data storage and management solutions such as Data Lakes, Data Lakehouse, Data Warehouses, Databases (DB), Distributed Databases (DDB), Federated Databases (FDB), and Multi-Databases (Multi-DB) in terms of sovereignty, interoperability, and trust. Table I presents a comparison of dataspaces with these related solutions. Notably, dataspaces secure data sovereignty by incorporating mechanisms for data owners to enforce access and usage policies, unlike Data Lakes that focus solely on the aggregation of vast amounts of raw data and often lack comprehensive governance features. Compared to Data Warehouses, which are tailored for structured data and analytics within a single organization, dataspaces enable dynamic data sharing across a myriad of stakeholders with stringent access controls and contractual agreements. Traditional databases primarily manage data within a single entity, while dataspaces provide a framework for secure and sovereign data sharing using standardized protocols to ensure interoperability. They support seamless and secure data exchanges across different domains and jurisdictions, a capability that Distributed Databases—focused on intra-organizational data distribution—do not inherently address. Federated Databases integrate multiple databases but often lack the standardized protocols and governance mechanisms inherent to dataspaces. Similarly, Multi-Databases combine databases but do not inherently tackle governance, sovereignty, and standardized interoperability challenges. Overall, dataspaces furnish a robust infrastructure for modern sovereign data management, enabling secure, controlled, and standardized data sharing across multiple stakeholders, thereby addressing the limitations of traditional data management solutions in a globally interconnected digital economy.

In the next part of this section, we will provide an insightful study of various initiatives (e.g., IDSA [1], Gaia-X/GXFS[2], DSSC[3], Eclipse Dataspace Components[4], CKAN[5], Fraunhofer FOKUS Piveau [2]) aimed at the development and operation of dataspaces by offering key technologies, frameworks, tools, and standards that facilitate secure, interoperable, and compliant data sharing and management. These collective efforts contribute to creating a robust infrastructure that supports data-driven innovation while ensuring data sovereignty and adherence to regulatory requirements.

[1]https://digital-strategy.ec.europa.eu/en/factpages/data-act-explained

[2]https://www.gxfs.eu/

[3]https://dssc.eu/space/BVE/357073105/Conceptual+Model+of+Data+Spaces

[4]https://projects.eclipse.org/projects/technology.edc

[5]https://ckan.org

TABLE II
COMPARATIVE ANALYSIS OF DATASPACE INITIATIVES

Initiative	Data Catalog	Data Connector	Trust Framework	Governance Model	Standardization	Open-Source / Proprietary	Effectiveness Comparison
IDSA	Manages metadata and data assets	Secure connectors for data exchange	Ensures data security and sovereignty	Governance policies and legal frameworks	IDS Reference Architecture	Open-Source	Effective in secure data sharing and compliance
BDVA	No specific catalog, focuses on analytics	Secure data sharing	Emphasizes ethics and privacy	Promotes standardization and ethical data use	Big Data Value Reference Model	Open-Source	Effective for data analytics and ethical data use
DSSC	Supports cataloging and integration	Facilitates secure data exchange	Trust and compliance framework	Compliance with data protection regulations	Follows European Data Strategy Guidelines	Proprietary	Effective for integration and interoperability
GXFS	Federated catalog for data discovery	Identity and trust services	Compliance and certification	Policy rules and certification processes	Gaia-X standards	Open-Source	Effective for federated services and compliance
Eclipse XFSC	Part of the core framework	Facilitates secure, interoperable exchange	Ensures security compliance	Policies for usage control and compliance	Eclipse Foundation Standards	Open-Source	Effective for building secure and interoperable data spaces
Piveau	Catalog and central data management component	Secure peer-to-peer connectors and integration/sharing	Maintains data sovereignty	Compliance with security and privacy standards	Adheres to open data standard	Open-Source	Effective for secure peer-to-peer data exchange and cataloging
CKAN	Open-source data management system	Limited, mainly data publication	Limited, focuses on metadata management	Adherence to data protection standards	Adheres to open data standards	Open-Source	Effective for data management and metadata handling
Advaneo	Comprehensive data catalog	Secure data transfers	Privacy-preserving data hub	Compliance with data sharing agreements	Follows European data strategy guidelines	Proprietary	Effective for data management and sharing

The International Data Spaces Association (IDSA) provides a Reference Architecture Model (RAM) for secure and trusted data sharing, alongside key components like Connectors, Metadata Brokers, Clearing Houses, and Identity Providers to ensure interoperability and data sovereignty, while establishing governance guidelines compliant with legal and ethical standards[6]. In parallel, the Data Spaces Support Centre (DSSC) offers federated services, infrastructure for integrating various data sources, collaboration tools, and a compliance framework for data protection and privacy [3]. The Big Data Value Association (BDVA) also supplies a reference model for designing dataspaces, tools for data analytics and secure sharing, and promotes standardization and ethical compliance [4]. Notably, Gaia-X Federation Services (GXFS) also maintains a federated catalog for data discovery, trust services for secure interactions, and compliance monitoring with Gaia-X standards [5]. While, Eclipse XFSC (Cross Federation Services Components) features a robust framework for dataspace management, ensures secure and interoperable data exchanges, and maintains compliance with security standards [6]. Besides those initiatives, Fraunhofer FOKUS also launched Piveau, which provides a central platform for dataset management, high-quality metadata management, and tools for data cataloging and secure sharing [2]. In parallel, CKAN offers an open-source data management system for dataset publication and sharing, enhancing data discovery through metadata management while at the same time, ensuring data protection compliance [7]. Interestingly another German-based SME Advaneo launched a dataspace management solution, which offers a data catalog and marketplace for easy discovery and monetization of data, providing a trusted data hub for privacy-preserved collaborative processing and secure data transfer [8]. In light of these initiatives, Table II presents a detailed analytical overview of the contributions and key outcomes of each initiative. This analytical comparison will aid us in identifying and adopting the appropriate initiatives and techniques necessary for developing the AC^3 DMPaaS.

[6]https://github.com/International-Data-Spaces-Association/IDS-RAM_4_0

III. THE AC^3 DATA MANAGEMENT PLATFORM AS A SERVICE

Data management within the Cloud-Edge-Client Continuum architecture represents a foundational pillar for AC^3, essential for the orchestration, optimization, and deployment of applications across the continuum. Our goal is to facilitate seamless, efficient, and secure interactions between diverse data sources and the continuum's applications and services, achieved by integrating advanced data handling, processing, and exposure functionalities with the application deployment and management processes. This innovative approach ensures that data is not only accessible but also actionable, within the overall service lifecycle and dynamically adapted to the end user's application-specific requirements, the application profile, and the infrastructure's status. The innovation of data management as an integrated PaaS simplifies data handling, making it easier for developers to work with data resources while creating a cohesive environment for managing computing, data, and network resources, fostering smoother integration and collaboration among stakeholders. At the core of this design are the AC^3 Data Catalog and Data Management Application Addons, which streamline the complexities associated with data access and handling. These components encompass a range of functionalities, including data indexing, retrieval, parsing, storing, and streaming from the data sources integrated within AC^3. They act as a bridge between application developers and diverse data sources, providing intuitive APIs while supporting multiple data formats and storage systems.

In the rest of this section, we will focus on each one of the 3 areas targeted by AC^3 in the path towards developing a true DMPaaS for the cloud edge continuum.

A. *Interfacing with End Users*

This area focuses on the mechanisms through which end users interact with the data managed within the AC^3 architecture. The **Data Catalog** provides a centralized platform for users to search, discover, and access available data assets. Simultaneously, data exposure functionalities ensure that users can securely and efficiently utilize data assets within their applications and services, in compliance with data governance and privacy policies. The goal in this domain is to enhance the user interface and experience of interacting with the data catalog to make data discovery and access more intuitive and efficient, strengthen data exposure mechanisms to support secure, compliant, and flexible data sharing, enabling users to leverage data assets across various applications and services.

Sharing data in the context of the AC^3 DMPaaS is based on the AC^3 Data Catalog service. This service is instrumental in providing the appropriate descriptions, metadata, information, and guidance to software developers regarding offered datasets and data sources. These data sources, incorporate a broad spectrum of domains and need to be properly described and annotated to be easily searched and discovered. The annotation of the data sources is done using rich semantic information based on well-established ontologies provided by the owners of the data during the initial phase of the data source registration. In this portal, users can browse and check the datasets shared, initially, by the Use Cases of AC^3 and, later on, any other datasets made available through our platform.

The adopted approach in AC^3 is based on the adoption of a well-established and widely used solution like Piveau, instead of re-inventing the wheel and providing yet another incomplete solution for our users. Piveau is a European data management ecosystem that provides components and tools to support the entire data processing chain from harvesting, aggregation, provision, and use, focusing on open standards and high interoperability. Our goal is to adapt Piveau to AC^3 needs and provide changes as possible extensions to the Piveau maintainers as an additional contribution of AC^3 to the European Open-Source software community increasing the value, reach, and potential of these tools. The AC^3 Data Catalog therefore is composed of the following internal components:

- The Hub UI is a frontend service through which users can interact with catalogs and datasets.
- The Hub Repository manages and syncs the triple store and the search index, providing a rich RESTful API for interacting with the catalog's items.
- The Hub Search is an interface on top of an RDF Triple Store and an Elasticsearch instance used for indexing the stored datasets and catalogs.
- A Virtuoso RDF Triple Store for storing all the semantic information of the system.
- An Elasticsearch instance for indexing the stored data and providing advanced search capabilities.
- An Identity Provider for handling all the authentication, authorization, and accounting operations to allow users of AC^3 to interact with the data catalog and the rest of the project's services.

As part of our initial work on representing the datasets of our 3 use cases on the AC^3 Data Catalog, we have defined the basic parts needed to represent data and datasets from their data sources. Datasets and data sources in the AC^3 Data Catalog are represented using Resource Description Framework (RDF) descriptions and a textual syntax for RDF called Turtle that allows an RDF graph to be completely written in a compact and natural text form, with abbreviations for common usage patterns and datatypes. Turtle provides levels of compatibility with the N-Triples format and the triple pattern syntax of the SPARQL W3C Recommendation[7]. Each document is split into two parts. The first part contains lines that define prefixes that make it easier to reference Internationalized Resource Identifiers (IRIs) in the serialization of the graph. The second part of the document is the actual triples that describe the RDF document with information on the dataset and data source they describe.

Similarly, we work on providing templated descriptions for the datasets that are to be used for each of the use cases of AC^3. These descriptions will need to be refined and enhanced as the project advances, to provide more comprehensive and accurate information that can be effectively integrated with the rest of the AC^3 platform. In them, we store fields regarding the data available, their meta information, the location the data refers to, their supplier, the period of data collection, the license provided, as well as the means of distribution the application developer needs to follow to get access to the data.

B. *Interfacing with Data Sources*

The first task for enabling data processing in AC^3 is the integration with data sources. This section provides insights into the methods used for it, the technologies involved, the software components used, and the planned development path.

[7]https://www.w3.org/TR/sparql11-query/

Data Connectors serve as vital links that facilitate the flow of data between the continuum and external data sources. These connectors are engineered to accommodate a diverse array of data types, formats, and communication protocols, ensuring that data can be ingested from and dispatched to a wide range of sources, including IoT devices, cloud repositories, and edge computing platforms. Our primary goal is to enhance the adaptability and performance of data connectors, ensuring they can efficiently manage increasing data variety and volume. This involves incorporating advanced data handling technologies, improving real-time data streaming capabilities, and ensuring robust error handling and data integrity checks. In the context of AC^3, we focus on providing connectors for streaming IoT data, streaming video contents, and previously stored video and astronomical observation files based on the use cases of the project. These data types represent a wide range of data sources and can provide us with a great understanding of what different data types need in terms of implementation and a great pathway on how to provide additional implementations for other users in the future.

The main component that takes the role of providing access to the integrated data sources is called a **Data Connector**. The Data Connector is not a single piece of software but a concept that facilitates all the interactions needed to register the data available to the AC^3 Data Catalog and to serve the available data to applications. The registration of data sources starts from the AC^3 Data Catalog. This basic information provides the data source with a dedicated identifier for each dataset made available to application developers. This identifier is used by the Data Connector to provide updates to the dataset's description through the Hub Repo API. The API simplifies the management of datasets and integrates seamlessly with Virtuoso, making dataset management more accessible while promoting transparency, knowledge sharing, and collaboration.

The next step in this path is the implementation of the **Data Connectors**. These components are based on the well-established Eclipse Data Space Connector (EDC), designed as an open-source framework aimed to facilitate the controlled and secure sharing of data in decentralized environments. EDC provides a suite of functionalities that support the handling and distribution of IoT and video data—distinguishing between different data temperature states like hot and cold and addressing the varying processing needs for these data types. Amongst others, the EDC offers the following functionalities:

- **Real-time Data Streaming:** EDC supports real-time management, data ingestion and processing of streaming data from IoT devices and video feeds.
- **Efficient Data Routing:** Data streams are dynamically routed based on policies, client requirements, and data types, optimizing for latency and bandwidth.
- **Hot and Cold Data Management:** For 'hot' data, which requires immediate analysis for time-sensitive decisions, EDC facilitates fast access and low-latency processing. 'Cold' data, not frequently accessed, is managed differently, optimizing storage techniques and cost efficiency and using more cost-effective, scalable storage solutions.
- **Secure Data Sharing:** The framework ensures that data sharing adheres to specified governance policies and compliance requirements, crucial for sensitive IoT and video data. EDC incorporates sophisticated user and data access controls, encryption, and secure data transmission protocols to protect the integrity and confidentiality of data as it moves across different jurisdictions or within various sectors of the industry.

Our work leverages EDC's capabilities to build a robust data space on top of cloud infrastructures. Cloud infrastructures offer reliability, scalability, and performance, which are essential for effectively implementing a data space that handles large-scale, high-speed data streams and facilitates complex transactions and interactions between diverse participants. Additionally, they provide scalable cloud infrastructure that can dynamically support the intensive demands of streaming, processing, and storing both hot and cold IoT and video data. The deployment of our solution there enhances operational efficiencies, reduces latency, and improves data throughput. In summary, the use of EDC in conjunction with the Hub within the AC^3 is poised to create an advanced, secure, and efficient data space that enables effective data sharing and utilization, drives innovation, and fosters collaboration. As a result, based on the use of the EDC we expect to provide AC^3 end users and infrastructure owners with the following three connectors designed as extensions to the EDC:

- **Cold Data Connector:** The Cold Data Connector, based on the IONOS S3 EDC extension [8], is a pivotal component designed to integrate seamlessly with the Eclipse Data Space Connector (EDC), specifically leveraging the robust and scalable storage capabilities of cloud object storage services. Given the AC^3 project's focus on developing a secure and efficient data space, this extension plays a critical role by enhancing the functionality of the minimum viable data space setup offering the following key features:
  - Data Storage and Management: the extension allows for efficient storage management practices such as data replication and lifecycle policies, which help manage data through its various stages effectively.
  - Scalability: storage resources scale based on real-time demand without compromising performance or data availability, maintaining robust data operations.
  - Integration Simplicity: integrate smoothly with minimal configuration, facilitating ease of setup and maintenance, ensuring focus remains on data utilization rather than infrastructure management.
  - Encryption: data stored is encrypted at rest, protecting sensitive data from unauthorized access, aligned with stringent security.
  - Access Control: sophisticated access control mechanisms that comply with the data governance policies including permissions for data creation, modification,

[8] https://github.com/ionos-cloud/edc-ionos-s3

and deletion, which are crucial for maintaining data integrity and adherence to regulatory standards.
  - Audit Trails: monitoring and logging help create a transparent and traceable data handling environment pivotal for audits and compliance.
  - Data Resilience: high durability of stored objects with safeguards against data loss for building a reliable data space with data continuity and availability.
- **Streaming IoT Data Connector:** The Streaming IoT Data Connector covers a large type of data sources to be integrated into AC^3, including various time-series-based data sources. It is a pivotal component designed to ease the introduction of real-time data for applications that need to process information generated from various IoT and sensing infrastructures, like building monitoring, utility metering, power grids, Industrial IoT deployments, and others. A key functionality of this Connector is the ability to distribute the collected IoT data in real-time to client applications, providing scalability, ease of integration, and data security.
- **Streaming Video Data Connector:** The Streaming Video Data Connector covers specific data source connectivity needs in which live video data is collected from deployed on-site cameras, aggregated in the designated data space, and shared through the deployed application components. The 2nd use case of AC^3 is specifically designed to include this option. The connector design follows a similar approach to the IoT Data Connector with tighter latency and bandwidth requirements to ensure that the video feeds are delivered as fast as needed to their consumer applications. A key functionality of this is the ability to distribute the collected video feed data in real-time to client applications providing scalability, ease of integration, data security, and user privacy.

C. *Using Data Sources through Service Deployment*

The final step toward building data processing applications in the context of AC^3 is to include the actual data retrieval and processing components in the AC^3 application's lifecycle. This includes both the orchestration of the application, and its deployment and migration from Cloud to Edge and back, as needed. This section provides insights into the methodology used to achieve this, the technologies involved, the software components that will be implemented and used, and the planned improvements for the next period of the project.

This area emphasizes the utilization of integrated data sources to drive the deployment and ongoing management of services within the AC^3 architecture. By leveraging real-time and historical data insights, services can be dynamically adapted to meet changing user needs, optimize resource usage, and enhance overall service performance. The main target is to develop capabilities for dynamic service deployment that utilize integrated data insights for real-time adaptation and optimization. Enhance service management frameworks to incorporate data-driven decision-making processes, enabling continuous service improvement and ensuring optimal alignment with user requirements and system constraints.

Data processing in AC^3 is implemented as part of the deployed applications. For this purpose, we provide application developers with specific application add-ons and guidelines for developing their own application add-ons that can connect to the available data sources, consume the data available, and generate new data. The data processing application add-ons are expected in the form of deployable micro-services that provide target-specific extensions to applications. In this direction, the existence of 3 very different use cases allows us to investigate different requirements and data settings that will help us design our data processing components in a way that is best fitted to as many data sources and application needs. As all AC^3 applications are based on a micro-service architecture, our add-ons operate following the same logic, providing multiple microservices that can be merged to provide application developers with all the building blocks they would need.

For the design of data processing micro-services, we follow two main approaches: the Hot data approach for live data sources, and the Cold data approach for datasets stored before usage. The general approach is showcased in Figure 1 and represents the generic design of the data processing application add-ons. As part of the AC^3 application, the AC^3 CECCM can deploy the following components to allow for the processing of data in the context of the application: The Data Connector, a Message Broker, a Data Mapper, a shared storage location, and one or more Data Manipulators. The technologies used and the role of each add-on will be presented for Hot and Cold data settings in the rest of this section.

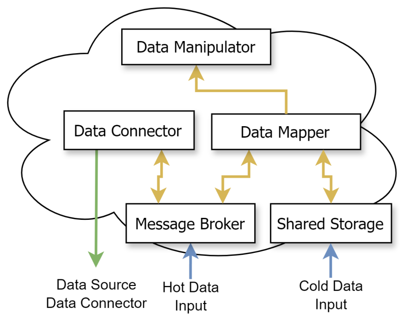

Fig. 1. AC^3 Data Management Application Add-ons

1) Adopted technologies: A core technology that is employed to allow for the transfer of data from data sources to data consumers in AC^3 is again the Data Connector. This software component assumes the responsibility of interfacing the Data Connectors on the data source's side to control the data retrieval for both the Hot and Cold settings. Secondly, tools for streaming and transferring the collected data are used, including Message Brokers like RabbitMQ and Kafka, using protocols like the Advanced Message Queuing Protocol (AMQP) to allow for secure, asynchronous, and reliable data transfer, regardless of the actual size of the data in question. Finally, due to the wide variety of applications that need to be deployed (visible through the AC^3 use case needs, but also considering any generic deployment scenario), specific end-user-defined application components will be used for the actual data processing. However, a common design feature in all these application components should be their ability to consume the retrieved data source data, as well as the need to provide the necessary feedback to the AC^3 platform for its operation. The implementation of the data management add-ons offers this common pool of components to service developers (or potentially providers) to adapt their processing application components to the data management application add-ons and successfully incorporate them in their design.

2) Hot Data: In a **Hot Data** setting, the AC^3 application employs a structured approach utilized for both two of the three project use cases, where each scenario involves the streaming of data to their respective applications.

The process begins with the *Data Connector* tasked with initiating data retrieval. Based on the application's specifications and configuration parameters, it interacts with the data source's counterpart to start the data streaming to the Application Message Broker. This broker acts as a crucial communication hub, ensuring data generated at the source is accurately transmitted to the AC^3 application via the hot data streaming interface, based on the AMQP protocol, with alternative protocols to be implemented as needed.

Next, the *Data Mapper* receives the incoming data via the Application Message Broker. Its primary function is to transform the data into a format comprehensible to the application's internal components, the *Data Manipulators*. In addition to format conversion, the Data Mapper may also perform minor processing or aggregation tasks, such as time-based analytics or data aggregation. Lastly, the *Data Manipulator*, which is integral to the application's business logic, integrates the data into the application's operational workflows (e.g., ML analysis, video processing, big data computations). The data manipulator is essentially the interface with the core application that uses the data to perform its business tasks.

3) Cold Data: In a **Cold Data** setting the AC^3 application is expected a bit simpler using less application add-ons.

The Cold Data retrieval begins again with the *Data Connector*. Acting on instructions defined within the application's deployment parameters, it interfaces with the data source's counterpart to initiate and orchestrate the transfer of the cold data to the *Shared Storage* of the application. Its interface, serving as a communication bridge, ensures that the data retrieval requests are accurately relayed to the data source's data connector. These repositories, where cold data resides, are accessed to fetch the required datasets. Then, the Cold Data Transfer Interface oversees the secure and efficient transfer of data from the repositories to the application's domain, ensuring data integrity and confidentiality. Upon successful transfer, cold data is temporarily staged in *Shared Storage*, a location accessible by the application's components to which the *Data Manipulator* has access. The Data Manipulator, dependent on the application's business logic, is poised for integration of the data into the application's operational workflows.

With the cold data being processed and ready, the add-ons are integrated into the application, enriching the application's functional capabilities and enhancing its decision-making processes. This integration marks the culmination of the cold data handling process, where the once dormant data now actively contributes to the application's dynamic environment, driving insights, and fostering informed decisions. Throughout this process, monitoring mechanisms are in place to ensure the smooth flow of operations and to address any potential issues promptly. The orchestration of these components and stages within the AC^3 architecture exemplifies a holistic approach to cold data management, ensuring that cold data is not merely stored but actively adding value to the application ecosystem.

IV. Conclusions

The AC^3 Data Management Platform as a Service (DM-PaaS) addresses the growing need for sophisticated data management solutions in the context of IoT and distributed computing environments. By integrating a dynamic data catalog, efficient data connectors, and a hybrid cloud-edge processing framework, the platform offers robust, scalable, and flexible data management capabilities. The platform enhances data discoverability, accessibility, and real-time processing, ensuring data integrity and consistency across the cloud-edge continuum. The use of standardized protocols and advanced metadata management techniques further supports seamless data operations and analytics. Through the comprehensive evaluation of related work and a detailed analysis of our implementation, this paper highlights the strategic necessity and advantages of such an integrated data management approach. The AC^3 DMPaaS not only facilitates efficient data movement and processing but also adheres to stringent data governance and privacy policies, making it a valuable solution for enterprises aiming to harness IoT data. Future work will focus on expanding the platform's capabilities, enhancing its adaptability to various data sources and applications, and continuing to contribute to the open-source community to foster innovation and collaboration.

References

[1] S. Steinbuss, B. Otto, A. Teuscher, and H. Pettenpohl, "Ids reference architecture model 4 (ids-ram 4)," 2022.

[2] F. Kirstein, K. Stefanidis, B. Dittwald, S. Dutkowski, S. Urbanek, and M. Hauswirth, *Piveau: A Large-Scale Open Data Management Platform Based on Semantic Web Technologies*, 05 2020, pp. 648–664.

[3] Data Spaces Support Centre (DSSC), "Starter Kit for Data Space Designers," https://dssc.eu/space/SK/29523973, 2023.

[4] E. e. a. Curry, "Data sharing spaces: The bdva perspective," in *Designing Data Spaces: The Ecosystem Approach to Competitive Advantage*. Springer International Publishing Cham, 2022.

[5] V. Siska, V. Karagiannis, M. Drobics *et al.*, "Building a dataspace: Technical overview," *Gaia-X Hub Austria*, 2023.

[6] Eclipse Foundation, "Eclipse dataspace connector," https://projects.eclipse.org/projects/technology.xfsc, accessed: 2024-07-07.

[7] S. Neumaier, J. Umbrich, and A. Polleres, "Challenges of mapping current ckan metadata to dcat," *W3C Smart Descriptions & Smarter Vocabularies (SDSVoc)*, 2016.

[8] Advaneo GmbH, "Data spaces solution," 2023, accessed: 2024-07-07. [Online]. Available: https://www.advaneo.de/en/data-spaces-solution-en/