10 DS2 Data Detection and Transformation Module (DDT)

10.1 DS2 Data Detection and Transformation Module (DDT)

Owner(s): SWAG
DOA Task: T4.2 and T4.3
Tier: 2
Nature: Optional
Results: K4.2 (Data Detection), Outcome (Transformation)

DOA Task: T4.3
This task will research and classify different forms of anomalous behaviour / error conditions over complex data lifecycles and create classifiers to recognise them. It will investigate the most appropriate AI algorithms, Methodologies to mitigate the effect of detected degradation of data quality will be investigated, and work together with the Human-centric Tools work package when human intervention is required.

DOA Task: T4.3
This task will prepare data to be exchanged between data requestors and data providers in disparate sectors and enterprises. Amongst the topics to achieve this goal, this task will cover the curation of generated/collected data, portability, and transformation of data, covering security data transport protocols and data interoperability issues, include transformations due to data sovereignty requirements, ontological reconciliation, and schema differences. Novel challenges to be addressed by this task include reconciling semantic differences in data classifications between sectors.

10.1.1 Introduction

Purpose: Dataspaces allow for data to be shared between data providers and data consumers. A lot of data comes from sensors and devices at a high rate. To allow for a well-defined data structure and quality during the data generation and exchange, DDT is a module that can analyse data on the fly.

Description: As real-time data analyser DDT will use the Software AG product Apama which is an event processing platform. It monitors rapidly moving event streams, detects and analyses important events and patterns of events, and immediately acts on events of interest according to your specifications. EPL, Apama's native event programming language, lets developers define rules for processing complex events. Such rules let the correlator find temporal and causal relationships among events. Apama can be detected to any event data source, database, messaging infrastructure, or application. In addition, the Apama programming environment contains classes used to define EPL plug-ins written in Python that run inside the Apama correlator. This is especially interesting since most data scientists developed there AI algorithms in Python and in this way, they can directly be integrated in Apama’s runtime environment.

10.1.2 Where this component fits

10.1.2.1 Big Picture

The data detection and transformation module, name DTT in the picture below, can be used by a participant and between participants who themselves could be within the same dataspaces or different ones.

DDT addresses the topic AI Detection (T4.2 Data Curation via Transformation (T4.3)

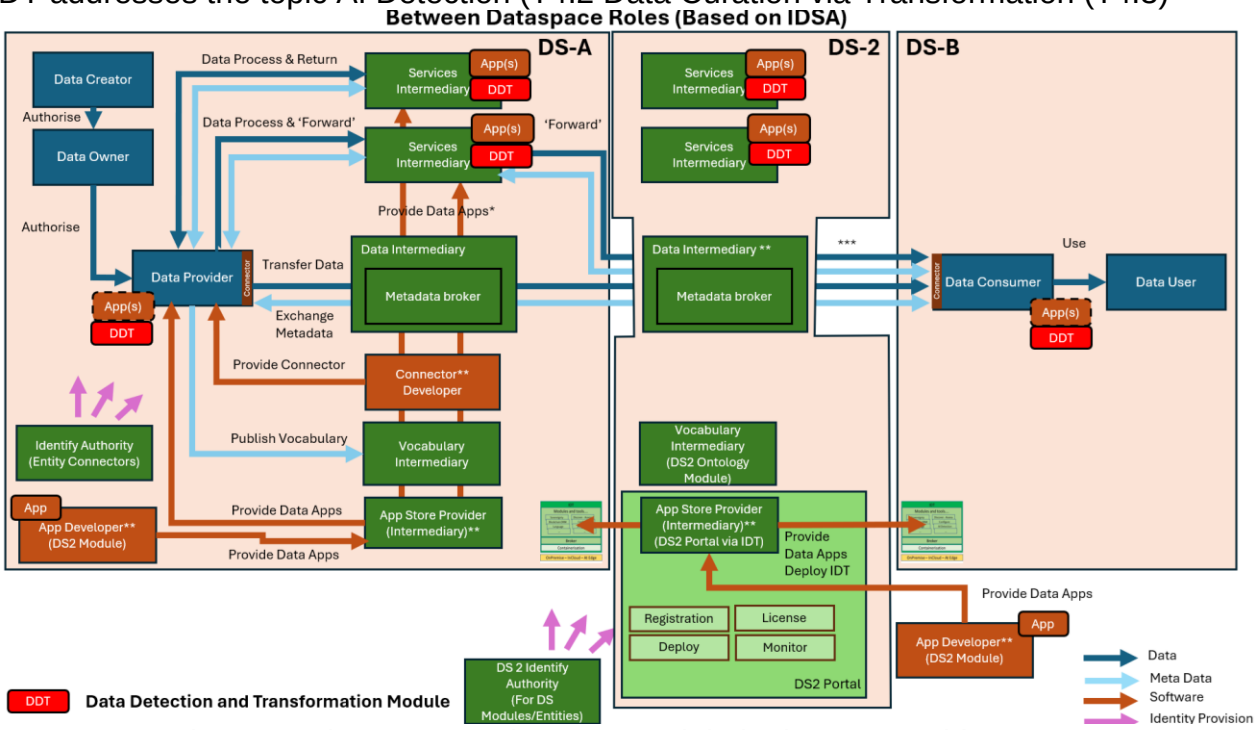


Figure 1: Placement of the DTT Module in the DS2 architecture

Where	Status
Within a single Dataspace for use between participants in that Dataspace only	Yes.
Deployed and used by a single participant to enable the participant in either an In-Data space or Inter- Data space scenario	Yes. The DTT module may operate at different places, e.g. at the data provider and/or data consumer.
Across Dataspaces without Service Intermediary	Yes. The DTT module may operate at different places in different dataspaces.
Across Dataspace with Intermediary	Yes.
Other Comments	N/A

10.1.2.2 Within a single Dataspace (where applicable)

DDT could be deployed by a data intermediary within a dataspace to provide a service for all data providers and consumers to check for anomalous behaviour and error conditions.

10.1.2.3 Deployed and used by a single participant (where applicable)

The data detection and transformation module can be deployed by the data provider or data consumer to constantly monitor the data quality and react if a degradation occurs. The transformation part can be used to correct the data if a degradation was detected or to apply custom transformations. The AI detection and transformation components can communicate with each other via their MQTT brokers if not run in the same instance.

10.1.2.4 Across Dataspaces without intermediary (where applicable)

The DDT can be deployed by the data provider in one dataspace and collect information on the data quality. This can then be passed on to the data consumer in another dataspace to use this information in the DDT transformation component.

10.1.2.5 Across Dataspace with Intermediary (where applicable)

The DDT can also be deployed by the data intermediary, it should be accompanied by a monitoring or logging component so that the fulfilment of the agreed policy on the data quality can be guaranteed and later proven in case of doubts.

10.1.3 Component Definition

The figure below represents the actors, internal structure, primary sub-components, primary DS2 module interfaces, and primary other interfaces of the module.

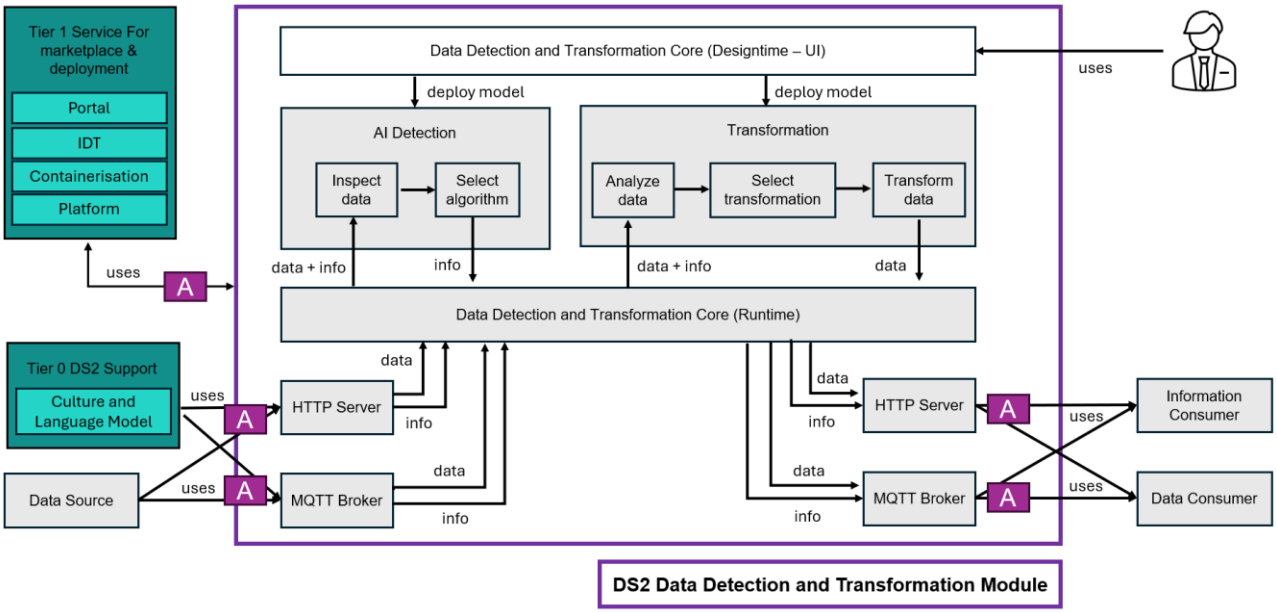


Figure 2: Schema for the Module

This module has the following subcomponent and other functions:

Data Detection and Transformation Module:

- **Data Detection and Transformation Core (Design time):**
  - This is an Eclipse-based UI with which:
    - the AI algorithms are developed, tested and deployed to the AI detection component

        - the transformation algorithms are developed, tested and deployed to the transformation component.

- **Data Detection and Transformation Core (Runtime):**
    - This is the heart of the Data Detection and Transformation Module. It is responsible for collecting the data input from the external source and pass it on to the AI detection and the transformation components. This uses either the HTTP or the MQTT APIs. In a similar fashion it also collects the information provided by the Cultural and Language module and passes it on to the other components. Once processed and used, the data and results of the data analysis will be sent out to the Information Consumer and the Data Consumer, also via the same HTTP or the MQTT APIs.
- **AI Detection:**
    - This component supports the execution of AI models developed using Python or other programming languages supported by Apama. It operates on data and returns information on the data worked on like if an anomaly has occurred or if the data quality has decreased. Many models will be provided to the user and the model selection is performed together with the T4.2 project partners INDRA and DIGI. It has subcomponents Data Inspection and Model Selection where based on the result of the Data Inspection, the appropriate AI model is selected.
- **AI Transformation:**
    - This component operates on the data passed on by the Core (Runtime) directly. It will detect schema changes and can transform the data into the correct format on the fly. It can also use the information provided by the DS2 Culture and Language module to define rules and limits what the data values should be and act accordingly, e.g. by eliminating out-of-range values. It has the subcomponents Analyze Data, Select Transformation and Transform Data.
- **APIs:** The Core runtime component has the following interfaces:
    - **HTTP Server:** This module will provide a HTTP interface to send and receive data and information from the DS2 Culture and Language module. Different URLs will be specified to distinguish between input and output channels and if the AI Detection or Transformation component should be used.
    - **MQTT Broker:** This is the de-facto standard for machine-to-machine communicate so an interface is provided to send and receive data from this module via publish and subscribe methods to certain topics. This can be used to specify the input and output channels of this module and also if the AI Detection or Transformation component should be used

External Components Used:

- **Data Source:** The data provider or data consumer, depending on where this module is deployed, configures and selects where the data to be analysed and transformed comes from
- **Tier 1 Service Stack for Marketplace and Development:** The module uses the portal to publish its configuration
- **Tier 0 DS2 Support:** The information on the data such as format and schema comes from the Cultural and Language module

- **Data Consumer:** The data provider or data consumer, depending on where this module is deployed, configures where the transformed data is sent to.
- **Information Consumer:** The data provider or data consumer, depending on where this module is deployed, configures where the information on anomalous behaviour, possible error conditions and the data quality is sent to.

External Interaction:

- **User:** The user uses the Eclipse-based UI to develop, test and deploy the AI detection algorithms and the transformation algorithms to the corresponding components.

10.1.4 Technical Foundations and Background

Apama was developed for fraud detection in the banking sector where a fast response to anomalies is needed to prevent misuse. The same principle can be applied to data in motion, e.g. generated by IoT devices or sensors and exchanged between different parties. With this module and its three provided sub-modules, degradation of data quality can easily be detected and mitigated via transformation or AI methods.

Subcomponent/Component	Owner	License
Apama Community Edition	SWAG	Freemium

10.1.5 Interaction of the Component

The following table specifies the primary input/output controls/data to blocks which are not part of the module

With Module/Feature	Receives From/Gives To	What
Cultural and Language Model	Receives From	Information on the data such as schema, frequency, units etc. coming from human interfaces
Data Source	Receives From	Data for Analysis
Information Consumer	Gives To	Data Quality KPIs
Data Consumer	Gives To	Data with improved quality

10.1.6 Technical Risks

Risk	Description	Contingency Plan
Changes to Data	Should the data change in form or function then the data model will need to be updated accordingly	Make sure the component is aware of changes to data. It will aim to raise a warning if the data has changed in format or schema
Inaccurate Results	The Data Detection and Transformation Module may not correct the data as wanted.	Test the algorithms thoroughly and implement additional controls.

10.1.7 Security

Security Issue	Description	Need
N/A	No special security issues are anticipated	

10.1.8 Data Governance

Data Governance Issue	Description	Need
Handling of personal data	This component is not set up to deal with the monitoring of personal data	User/Provider should ensure personal data transferred is transferred according to relevant regulations

10.1.9 Requirements and Functionality

This module will be used in the following use cases:

City Scape ✓

Green Deal ✓

Agriculture -
Inter-Sector TBD

Their requirements and functions/extensions to achieve them relative to this module, specifically extracted from the use case are as per the table below noting that in many cases further discussion might need to take placed between pilot partner and module partner to determine if a fit or the scope of the precise fit.

10.1.9.1 Data Detection and Transformation Module

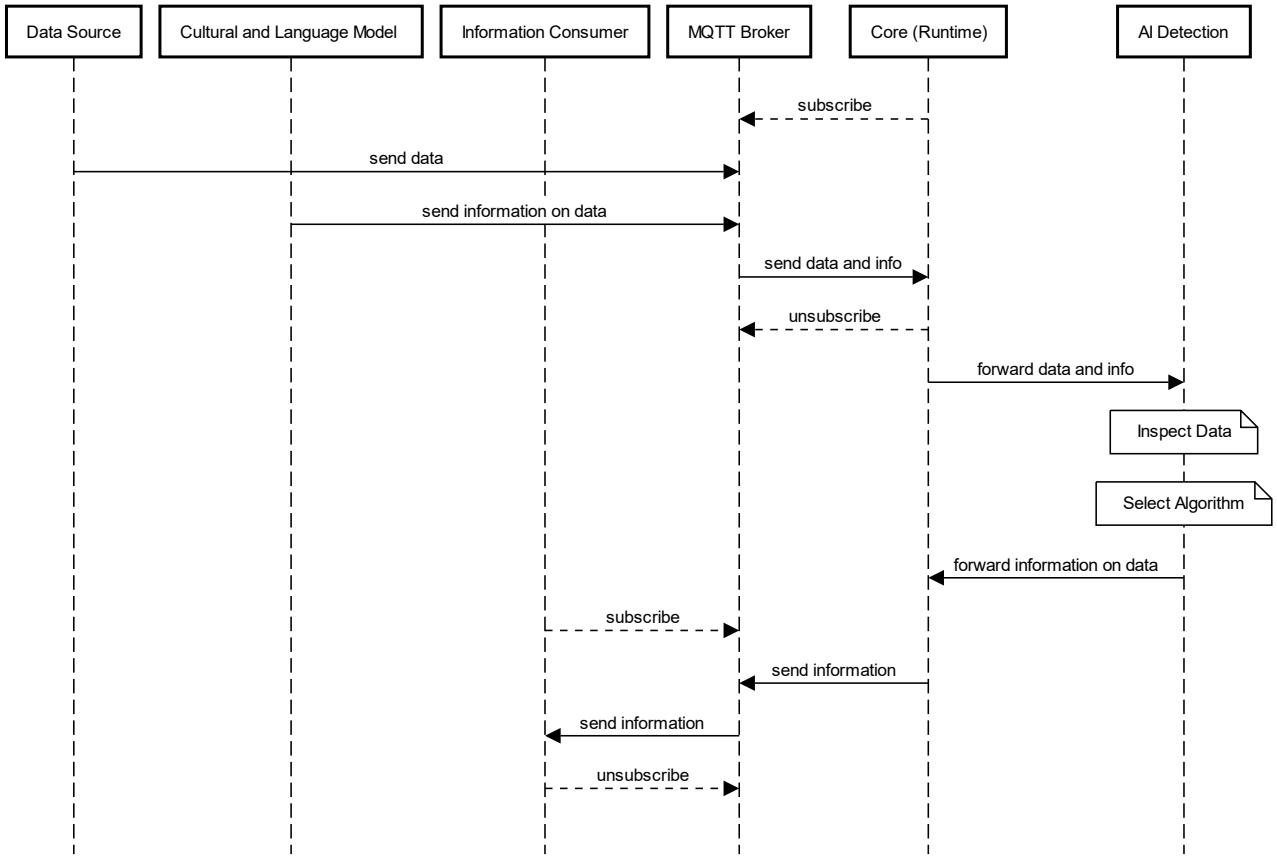
WHERE	WHAT	WHY	Run/Design Time	Priority
	Use Case 1: City Scape			
1.1.3.1	Requirements on the data: data quality, data provenance, data logging, and data harmonization	Right now, there is no integrative technology of the different sets of data and a lack of recommenders that are comprehensive of the different data types/sectors that influence the net zero quality of the city.	R & D	C
	Use Case 2: Green Deal			
1.1.3.2	“From the perspective of the data quality, we could develop some tools for	We need to know more details about the use case	R & D	C

	data harmonization … and transformation.”	for such tools and what the goals are.		
1.1.3.2	There are requirements for data storage security and data quality	primarily focused on handling publicly available data	R & D	
UC2.4 and UC2.5	Historical data can be used; refresh time is not important, but the data quality must be assured with respect of time and location accuracy.	MOMS is using the data from two sensors in the city. This data is shown on Municipality website, where citizens and other stakeholders can observe current level of pollution in the city	R & D	
2.2	Data Quality Module		R & D	M
	Use Case 3: Agriculture			
1.1.1	in order to involve actors from adjacent sectors and create sustainable business models, the need for data interoperability and portability		R & D	S
UC3.1, UC3.2, UC3.3	Data Quality Module		R & D	M
1.1.5	Data Quality Module Earth observation data / Satellite imagery	Satellite imagery provides a broad view of agricultural fields and can be used to monitor crop health, detect anomalies detecting changes in vegetation patterns, colour variations, and growth anomalies.	R&D	M

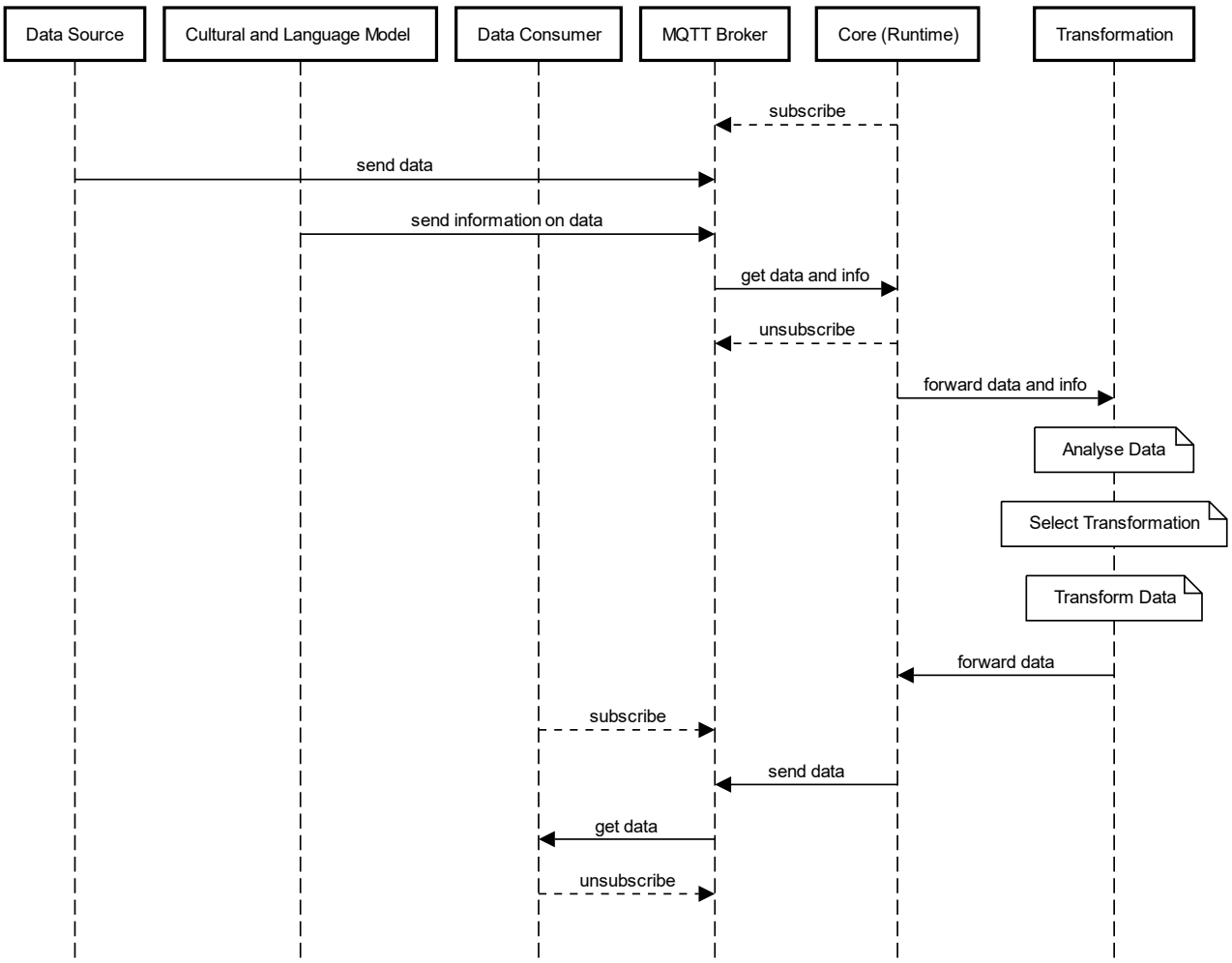
10.1.10 Workflows

The following sub-sections describe the sequence diagrams of the Module

10.1.10.1 Read Data Source, analyse it and check for anomalies and errors



10.1.10.2 Read Data Source, analyse it and apply needed transformations



10.1.11 Role, Resourcing, and Milestones

Sub-component	Main Activity	M18	M24	M30	M36
DTT Core Runtime	Initial Setup as Docker container				
DTT Core Design time	Initial Setup, test and document interplay with Core Runtime				
HTTP Server, MQTT Broker	Initial Setup as part of DTT Core Runtime, enable SSL/TSL communication				
Data Source, Data Consumer, Information Consumer	Connect sample external Date Source and Data Consumer				
AI Detection	Research suitable AI algorithms for anomaly and error detection and write corresponding Apama modules				
Transformation	Research suitable Transformation algorithms for portability and data interoperability and write corresponding Apama modules				
Culture and Language Model	Connect to DTT, retrieve necessary information and use it in AI Detection and Transformation sub-components				
AI Detection	Adapt AI algorithm to the pilots				
Transformation	Adapt AI algorithm to the pilots				
Data Detection and Transformation	Final integration of pilots				
Data Detection and Transformation	Documentation and final testing				
DS2 componentization, final integration					
Table Total/DOA Task Total/Resilience	**Comments: Effort of 3 PMs was shifted to T6.2 Data Quality Module**				

10.1.12 Open Issues

The following table summarise open issues/uncertainties that need to be resolved during the next stages or implementation.

Issue	Description	Next Steps	Lead or Related Component
Cultural and Language Model	Information on the expected data needs to be provided	Discuss data formats	INTU, SWAG
DTT	Discuss how information on the data can be exchanged between AI Detection and Transformation component if not deployed together	See if MQTT bridge can be used	SWAG
Discover, Trust	Do we need information from these modules	Check module description and discuss it with VTT	VTT, SWAG