Proactive and Preventive Data Governance

2021

Data sharing problems

Leads to

THIS

THIS

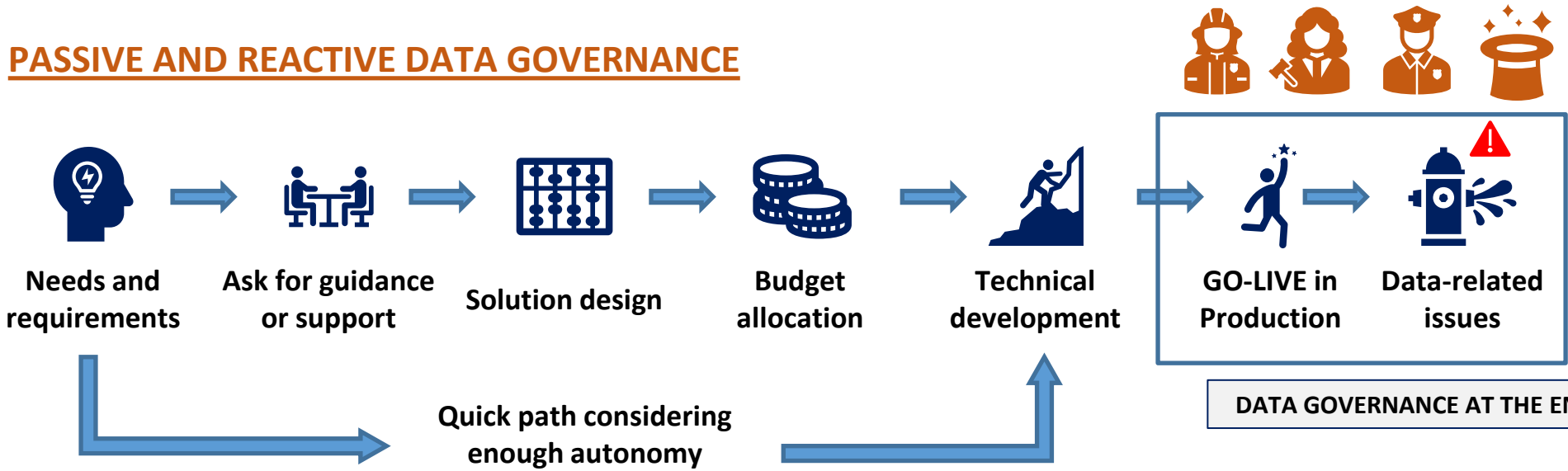
- Lack of awareness about where good-quality data is and who may have the corresponding knowledge
- Neither policies nor procedures for data sharing have been established and properly communicated
- Inefficient and uneffective ways to share data are the only ones available for users trying to share data
- Ignorance about which privacy or security rules needs to be applied to which data or which regulations must be met

- People do nothing because they do not know where to look for what they need or “The first they find, the first they use”
- People do not know how to share data so they don’t share it (no business value) or they share it in a bad way
- People share data just by sharing files through e-mail or messaging applications (thinking within the digital era)
- People do not know which data may be shared to whom and for what purposes so they don’t share it or regulations are not met

NO PROPER DATA GOVERNANCE

Different Data Governance approaches

PASSIVE AND REACTIVE DATA GOVERNANCE

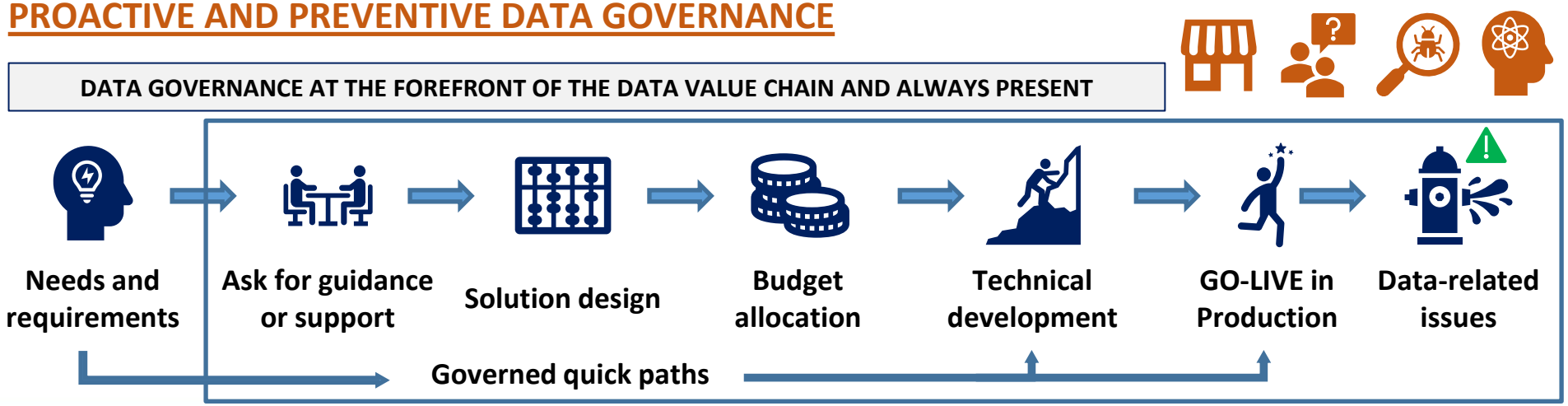


- × No Data Culture promotion (ownership, stewardship, accountability, …)
- × Project-based vision = Data Silos
- × No synergies = Inefficiencies
- × No data democratization/sharing
- × Errors by impacts in production
- × Regulatory & Risk-related issues
- × Data Quality & Security issues

DATA GOVERNANCE AT THE END OF THE DATA VALUE CHAIN

PROACTIVE AND PREVENTIVE DATA GOVERNANCE

DATA GOVERNANCE AT THE FOREFRONT OF THE DATA VALUE CHAIN AND ALWAYS PRESENT



- ✓ Data ownerhisp/stewardship from the beginning leads to high quality data
- ✓ Synergies and reuse = Costs reduction
- ✓ Integrated demand management
- ✓ Fostering data democratization/sharing
- ✓ Impacts management before production
- ✓ Governed self-service promotion
- ✓ Technical processes automation

Modern Data Ecosystem - Architecture

From a Data Governance perspective data sharing should be leveraged thanks to the establishment of the corresponding policies and procedures while ensuring data compliance and stewardship

We must know who is consuming which DATA and for what purposes

Data sharing needs must be satisfied in an **agile** way

Data access **policies and procedures** must be defined and followed

Data **stewardship, security and privacy** must be properly managed

We want to consume a lot of DATA for multiple purposes and needs

Data Governance

Governance model configuration
New Generation Business Glossary
Workflow Tools
Centralized metadata and lineage management
Active governance and dynamic audit
Data portal with shopping cart
Data Governance Dashboard
Data contracts for the use of information

Data producers

Data Services for Ingestion

aws, Azure, CLOUDERA, kafka, nifi, cdata, ETL ELT, Events Streaming, API

Identity & Access Management (Data Privacy & Security)

Sentry, Apache Ranger, SecuPi

Data Platforms (Storage, Modelling, Preparation, Processing, Publication & Sharing)

aws, Azure, hadoop HDFS, CLOUDERA, snowflake, Java JDBC, HIVE

Data Services for Consumption

aws, Azure, CLOUDERA, kafka, nifi, cdata, denodo, Data Tools, Java JDBC, API

Data consumers

DATA ASSETS

- Stored in various **repositories**
- Supported by different **technologies**
- Available in multiple **formats**

Modern Data Ecosystem - Governance

Data sharing needs must be satisfied in an **agile** way

Data access **policies and procedures** must be defined and followed

Data **stewardship, security and privacy** must be properly managed

PROACTIVE & PREVENTIVE Data Governance

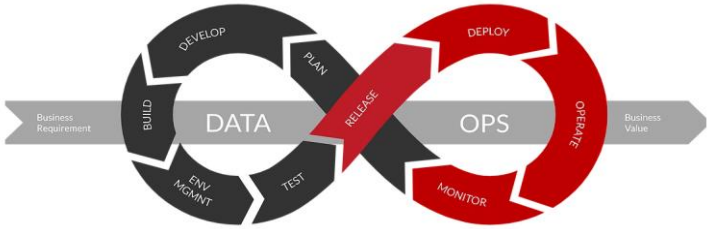
- Putting Data Governance at the forefront of **the data value chain**
- **Collaborative** approach empowering different stakeholders

- Integration with **demand management**
- **Integration** with other technologies and pieces
- **Automation** of common technical processes
- Incremental and iterative approach by **use cases**
- **Democratisation** and governed self-service
- **Monitoring** for continuous improvement

- **Metadata repository** at the heart of the data ecosystem
- Abstracting **data management** from technologies and platforms

Processes automation

- **METADATA EXTRACTION**
- **DATA SAMPLING**
- **DATA ACCESS MANAGEMENT**
- **DATA STRUCTURES MANAGEMENT**
- **EXTERNAL DATA PROCESSES AUDIT**
- **OTHER INTEGRATIONS**



DATAOPS NEEDS PROACTIVE AND PREVENTIVE DATA GOVERNANCE WITH PROCESSES AUTOMATION

Active governance

Integration over the Data Platforms for technical processes automation

Active governance is the ability of Anjana Data to automate common data management technical processes needed for data sharing over the Data Platforms thanks to its integration capabilities

DATA ACCESS MANAGEMENT

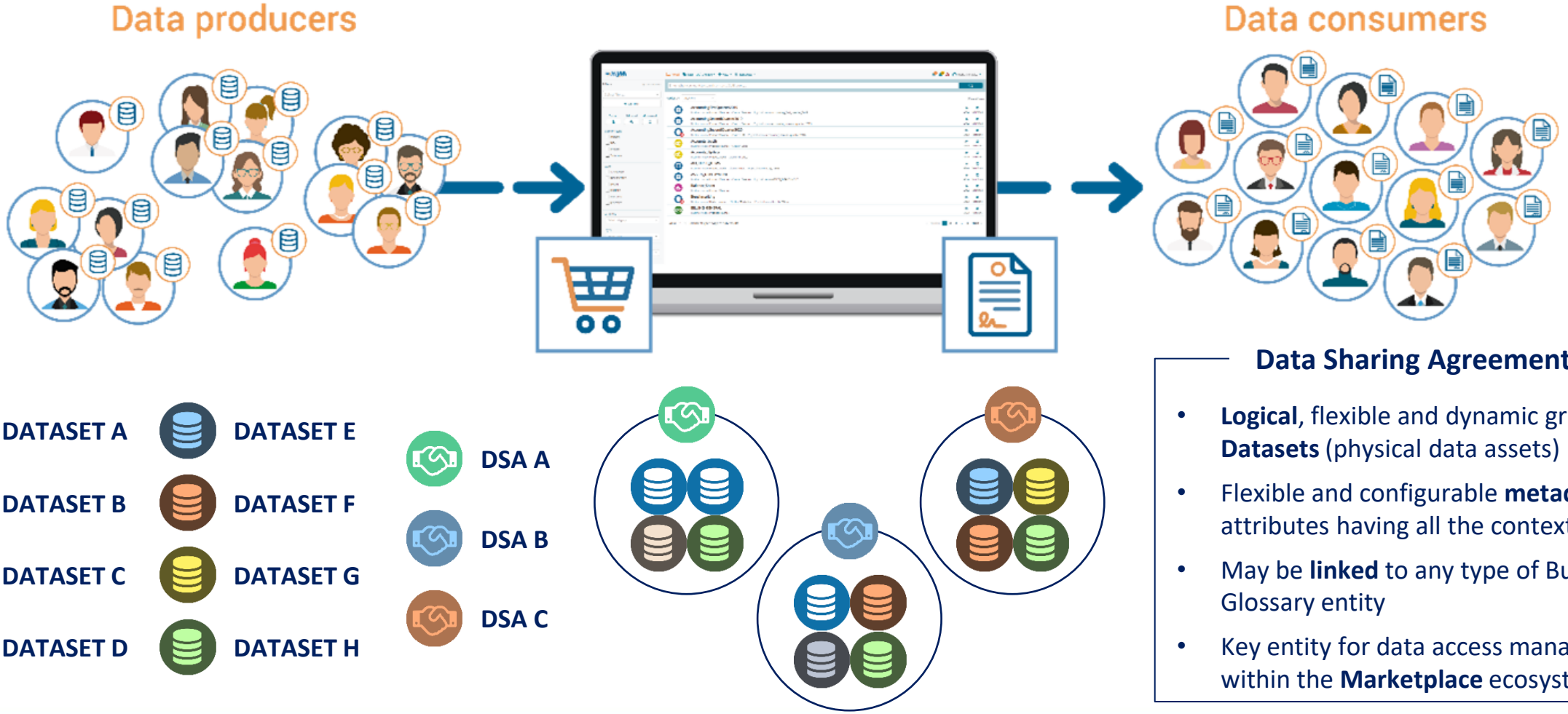
- Ability to automate the granting of privileges after the corresponding approval workflow
- It is done at **object level** (dataset, process/instance path) and **user level** (nominal or service)
- It implies on the one hand the connection with identity management (user) and on the other hand with privileges management (user profile, linked group, ...)
- In addition, Anjana will need to be connected to the corresponding LDAP/DA or identity management system for roles and users management

DATA STRUCTURES MANAGEMENT

- Ability to automate the creation, modification and deletion of data structures and points of storage (i.e. tables, views or paths)
- It is more thought for **NoSQL, Big Data and Cloud environments** where there is a culture of data exploitation with dynamic storage
- It is necessary to know how the metadata and the schemas are managed in the system and which actions would have to be launched from Anjana to manage the structures or storage systems

Active governance: Data access management

Data Portal and Marketplace. Operationalizing DSAs

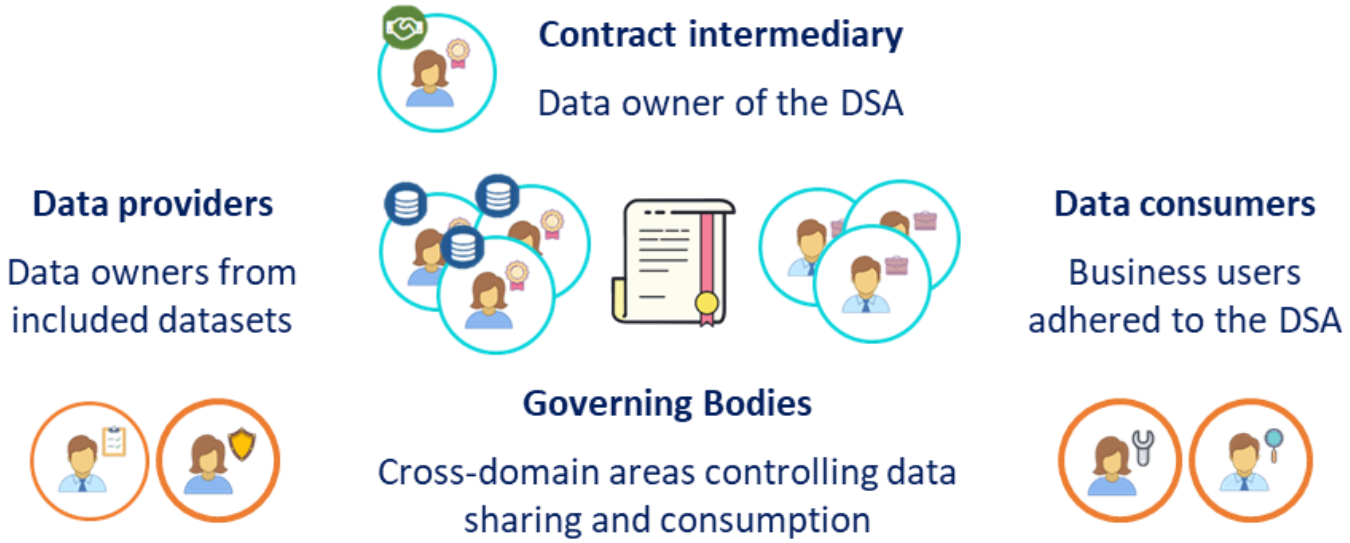


Data Sharing Agreements

- **Logical**, flexible and dynamic groups of **Datasets** (physical data assets)
- Flexible and configurable **metadata** attributes having all the context
- May be **linked** to any type of Business Glossary entity
- Key entity for data access management within the **Marketplace** ecosystem

Active governance: Data access management

DSAs Governance: Policies, Procedures and Data Contracts



Establish a common and standard model for the **management of data access** through the implementation of **multi-stakeholder Data Contracts** between producers and consumers

Like any **data asset,** they must have their own **policies** and **procedures** within the defined data governance model and according to the **data domain** to which they belong

They can involve several **stakeholders**, define different **categories**, identify various levels of data **sensitivity** and **confidentiality**, support **versioning** and have **expiration** date

Active governance: Data access management

DSAs Categories

ROADMAP V5.0

CATEGORY I

- They only grant **read permissions** to the Datasets included in the DSA
- Consumer users adhere to these DSAs in order to be able to consume the data of the Datasets included in the DSA with their **nominal user**
- **Example: an user willing to consume several tables**

CATEGORY II

- They only grant **read permissions** to the Datasets included in the DSA
- Consumer users adhere to these DSAs in order to be able to consume the data of the Datasets included in the DSA with a **sevice user**
- **Example: Report, Application, Digital use case, …**

CATEGORY III

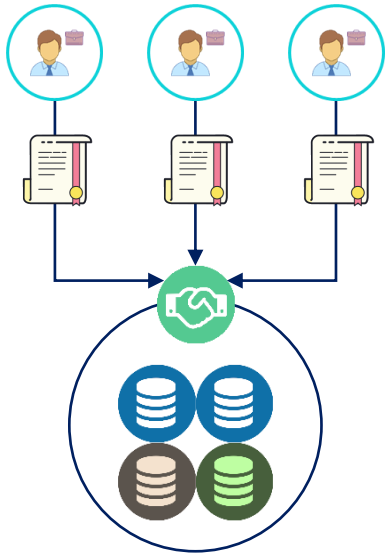
- They can grant **both read and write permissions** to the Datasets included in the DSA
- Consumer users adhered to these DSAs in order to get the defined permissions on the Datasets included in the DSA with their **nominal user**
- **Example: DataLab/SandBox for nominal user**

CATEGORY IV

- They can grant **both read and write permissions** to the Datasets included in the DSA
- Consumer users adhered to these DSAs in order to get the defined permissions on the Datasets included in the DSA with a **service user**
- **Example: DataLab/SandBox for service user**

Active governance: Data access management

Integration with Identity and Data Access Management Systems



Physical DSA

- A group in an Identity Management System (i.e. LDAP, Active Directory, IAM) containing users (nominal or service)
- A role in a Security and Privacy System (i.e. Sentry, Ranger, Active Directory, IAM) containing privileges (CRUD, RWX, policies) over Datasets (tables, files, …)

- Unambiguous identification of data **producers and consumers**
- **Data access policies** for specific use cases (encoding, encryption, masking, anonymization, filtering, ...)
- **Requirements for producers** for the correct consumption of the information contained in the Datasets (Quality, Availability, SLAs, ...)
- **Licensing terms and conditions of use for consumers** (associated costs, transferred rights and liabilities, additional clauses, ...)

Technical processes automation

Integration with Identity Management Systems and Technologies

- Each DSA is a group in the IMS
- Each time an user get adhered to a DSA, it will be included in the IMS group

Integration with Security and Privacy Systems and Technologies

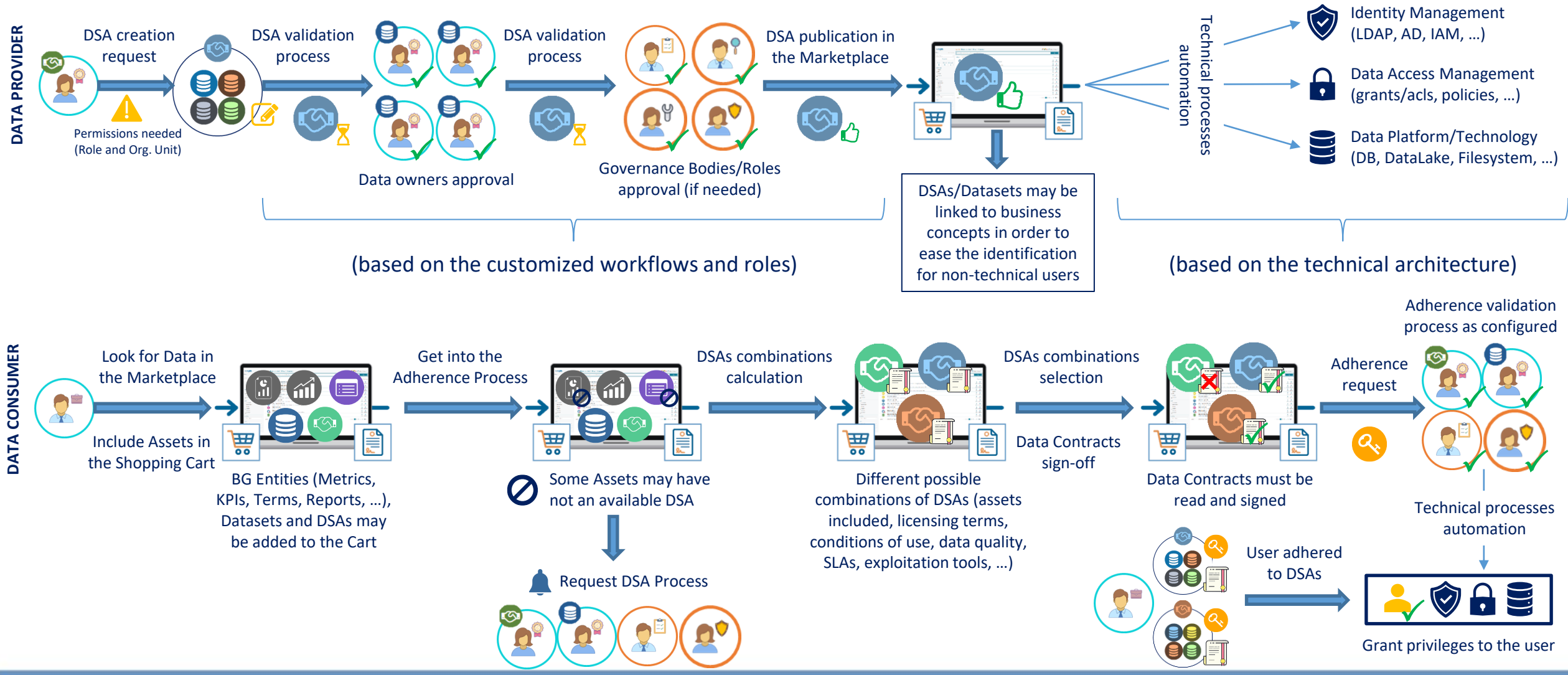
- Each DSA may be a role/group/profile in the corresponding SPS
- Each role/group/profile will have the corresponding grants/acls/policies over data

Integration with Data Platforms and Data Services for Consumption

- Each DSA will define the corresponding data access policies for the Datasets
- Each time a Dataset is included in the DSA the defined policies will be enforced

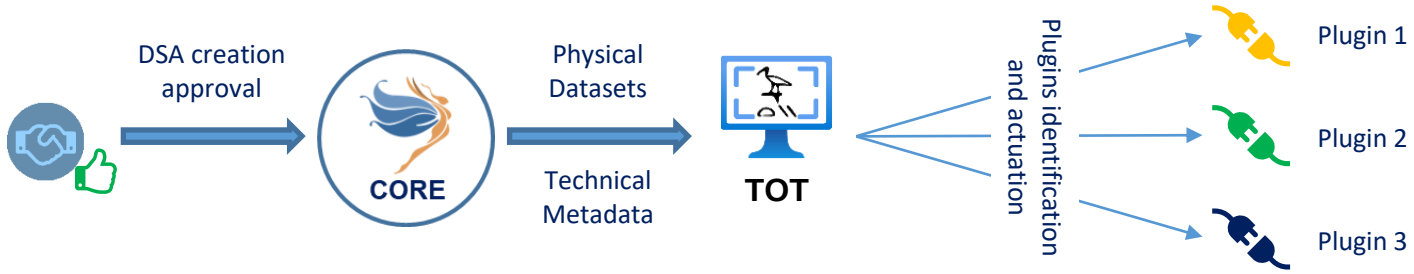
Active governance: Data access management

Integration with Identity and Data Access Management Systems



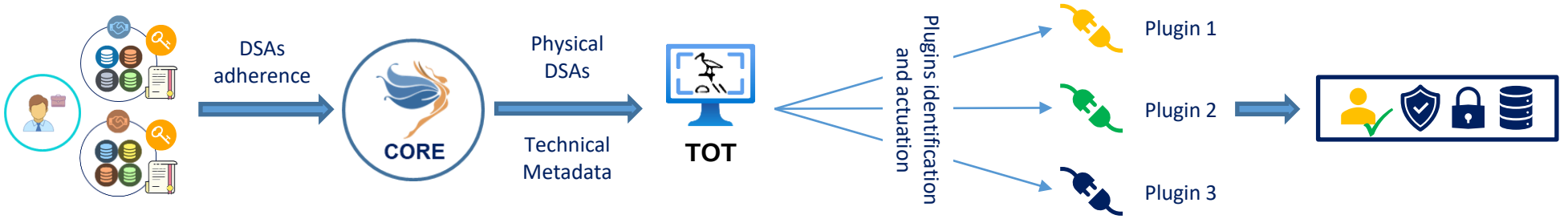
Active governance: Data access management

Integration with Identity and Data Access Management Systems



- ✓ **Each plugin works over a logical partitioning resolving a connection**
- ✓ **Each plugin may launch multiple actions based on the technologies implied**
- ✓ **A technology may have more than one plugin depending on its possible uses**

- Identity Management (LDAP, AD, IAM, …)
- Data Access Management (grants/acls, policies, …)
- Data Platform/Technology (DB, DataLake, Filesystem, …)



- ✓ **Users will have privileges to access physical data where stored/served**
- ✓ **Data is not accessed or served by Anjana Data but using other tools**
- ✓ **Complexity of applicable policies depend on the technologies behind**

SECURITY NEEDS

Depending on the actions to be performed by the plugins, the required point-to-point connectivity must be guaranteed (IPs, Ports, firewall, …) and the service user must have the required privileges and credentials over the corresponding technologies

EXTENSIBILITY CAPABILITIES

Available plugins may be configured or even modified to serve for specific purposes or ad-hoc implementations as well as new plugins may be developed and deployed following the recommended standards and practices provided by Anjana Data

Active governance: Data access management

Integration with Identity and Data Access Management Systems

The relation Dataset-DSA is established in the DSA creation/edition process

The permissions that apply to that Dataset because it is included in that DSA are indicated in the relation

ROADMAP V5.0

- (LDA
- Data Acces (grants/acls, po
- Data Platform/Technology (DB, DataLake, Filesystem, …)

The relation User-DSA is established in the DSA adherence process

The user (nominal or service) which will get the permissions is specified in the relation

PERMISSIONS RELATIONSHIPS

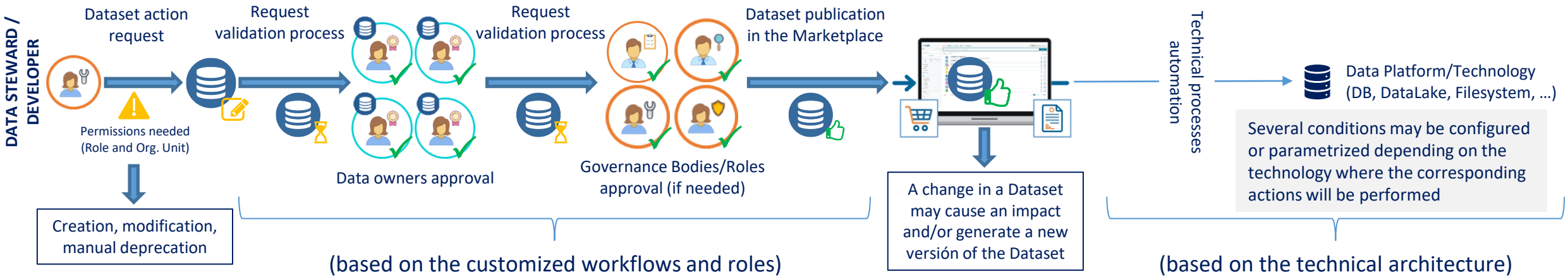
- Specific attributes can be configured to be requested to the user when relating Dataset and DSA such as a BG entity, an attached file, a text field or a date.
- The complexity of the permissions that can be applied (column/row, masking, ...) depends on the underlying technologies.

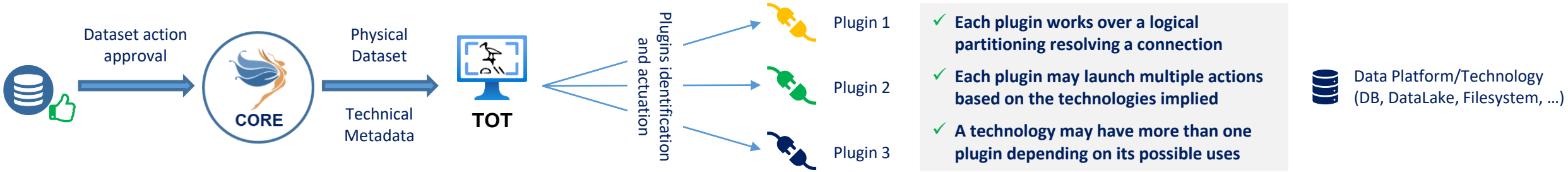
ADHERENCE RELATIONSHIPS

- Specific attributes can be configured to be requested to the user when requesting membership to a DSA, such as a BG entity, an attachment, a text field or a date.
- The users adhered to the DSAs (nominal or service) must already exist in the Identity Management System.

Active governance: Data structures management

Integration with Data Platforms





SECURITY NEEDS

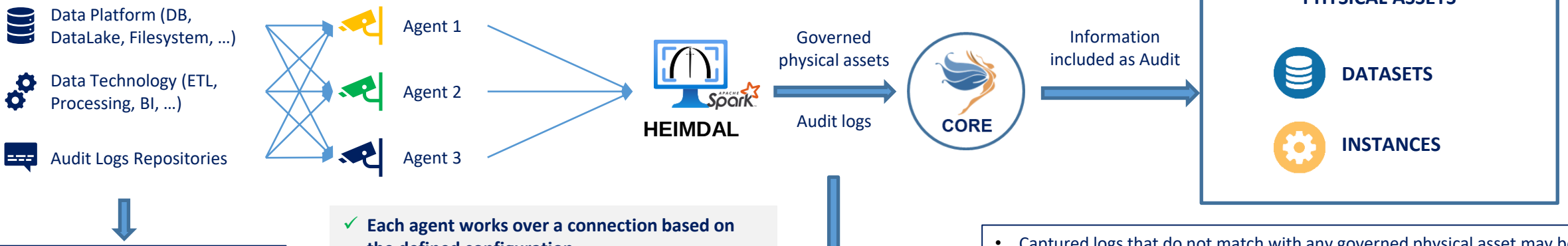
Depending on the actions to be performed by the plugins, the required point-to-point connectivity must be guaranteed (IPs, Ports, firewall, …) and the service user must have the required privileges and credentials over the corresponding technologies

EXTENSIBILITY CAPABILITIES

Available plugins may be configured or even modified to serve for specific purposes or ad-hoc implementations as well as new plugins may be developed and deployed following the recommended standards and practices provided by Anjana Data

External data processes audit

Monitorization of audit logs from Data Platforms and Technologies



- Processes executed over the Datasets in batch / real-time modes
- Queries or actions launched by the users over the Datasets
- Modifications over the Datasets
- Events generated by code

- ✓ **Each agent works over a connection based on the defined configuration**
- ✓ **Each agent may capture several actions depending on the involved technologies**
- ✓ **A technology may have more than one agent depending on its possible uses**
- ✓ **An agent may capture ad-hoc events generated by code if designed in that way**

- Captured logs that do not match with any governed physical asset may be analized for inconsistencies detection
- Alerts may be generated based on the interpretation of captured logs which have been included as Audit
- Inconsistencies detection may be performed thanks to the analysis of captured logs and declarative metadata (i.e. processes which their execution parameters do not match with its definition, unauthorized accesses by users not adhered to any DSA containing a specific Dataset, …)

SECURITY NEEDS

Depending on the actions to be performed by the plugins, the required point-to-point connectivity must be guaranteed (IPs, Ports, firewall, …) and the service user must have the required privileges and credentials over the corresponding technologies

EXTENSIBILITY CAPABILITIES

Available plugins may be configured or even modified to serve for specific purposes or ad-hoc implementations as well as new plugins may be developed and deployed following the recommended standards and practices provided by Anjana Data