DAP by Dateligens

Technical summary

Writer	State	Commentary
Jean-Michel Invernizzi	Approuvé	01/07/2023

DAP software : technical summary

1. Software Description

- **Overview:**

Dateligens Anonymization Platform (DAP) , offers a robust and innovative approach to anonymizing large volumes of data (big data), ensuring compliance with privacy regulations while maintaining data usefulness.

In today's data-driven world, organizations face significant challenges in protecting sensitive information while leveraging the power of big data analytics. Privacy regulations, such as the General Data Protection Regulation (GDPR) in Europe, impose strict requirements for data anonymization.

However, existing solutions often fall short in striking the right balance between privacy and data utility (internal or external processing). This presents a significant market opportunity for Dateligens to address the pressing need for an innovative and dynamic big data anonymization solution.

DAP solution can be used in SaaS mode or On-Premise.

DAP is intended for companies business users as well as IT departments, CISO (Chief Information Security Officer), CDO (Chief Data Officer) and DPO (Data Protection Officer).

- **Architecture:**

The architecture of DAP consists of several key components that work together to achieve dynamic anonymization of data. Let's explore these components:

- Data Ingestion: DAP supports the ingestion of diverse data sources, such as databases, data lakes, and streaming platforms. It can seamlessly integrate with these sources to fetch the data and prepare it for the anonymization process.
- Anonymization Engine: At the core of DAP lies the anonymization engine, called Pantheon. It’s a proprietary solution, responsible for applying various anonymization techniques to the data. It leverages the computational capabilities of Spark to handle distributed processing and scalability requirements. The engine employs a combination of generalization, suppression, perturbation, and other anonymization methods to protect sensitive information while maintaining data utility.
- Metadata Management: DAP incorporates a metadata management component that allows users to define and manage privacy rules and policies. It provides a user-friendly interface or API for specifying data elements to be anonymized, along with the desired anonymization techniques and parameters. The metadata management system ensures consistency and facilitates the application of anonymization rules across different datasets.
- Anonymization Algorithms: DAP encompasses a comprehensive set of anonymization algorithms tailored for different data types and privacy requirements. These algorithms are implemented using Scala, enabling flexibility and extensibility. DAP can provide a rich selection of techniques such

as k-anonymity, l-diversity, t-closeness, and differential privacy, enabling users to choose the most appropriate method for their specific use cases.

- Performance Optimization: Spark's distributed processing capabilities enable DAP to efficiently handle large volumes of data. DAP leverages Spark's optimizations, such as data partitioning, in-memory caching, and parallel execution, to maximize performance and minimize processing time. This allows for near-real-time anonymization of big data, even in high-throughput scenarios.
- Integration and Deployment: DAP provides integration capabilities with various data platforms, analytics tools, and data processing pipelines. It supports seamless integration with popular big data frameworks like Apache Hadoop and Apache Kafka. DAP can be deployed in both on-premises and cloud environments, allowing organizations to leverage their existing infrastructure and scale as needed.

2. Data Anonymization Techniques:

    - Description of Techniques:

The DAP platform allows to upload large volumes of data and then process them quickly in order to comply in particular with the RGPD / GDPR rules. DAP allows the choice between several processing modes.

Data can be transferred manually or automatically through the web (upload / download https, test platform accessible at: dap.dateligens.com), from your server, your database or as real time events supporting encryption (as tested in our long time POC with Société Générale Business Services).

The web portal allows you to create a customer account and manage various projects that can be launched manually (in test) or automatically (in production). Your transformed data can then be downloaded to your own infrastructures.

DAP is a scalable and industrial solution based on proprietary algorithms allowing parallel processing on servers clusters while respecting data time-series, particularly important in the context of IoT or cybersecurity (attack patterns) using Web / Spark / Scala technology.

Different techniques can be realized by DAP to reduce and control the data leakage risks with one unique anonymization engine:

- Anonymization : irreversible processing consisting in rendering impossible any identification of the person by any means whatsoever.
- Pseudonymization : tokenization processing of personal data to no longer assign the data to an individual without an additional information, consisting in replacing the directly identifiable data (name, for example) of a set of data with indirectly identifiable data (aliases , for example).
- Tokenization : a process which makes it possible to replace critical data (personal or not) with an equivalent data which will have no intrinsic value or exploitable meaning once it has left the

system. One unique particularity of DAP is to allow you to manipulate these tokenized data without significant risk and then, **if desired**, to de-tokenize them at the end of the process.

- K-anonymization (data blurring) : technique guaranteeing that a multi-criteria search, combining quasi- identifying attributes, can't isolate a person in a group of less than K individuals.

Different functions existing in DAP or in development:

Anonymization Manager: non-reversible anonymization unitary process.

Tokenization Manager: reversible anonymization unitary process and De- Tokenization Manager (data restoration).

Data Blurring Manager: data blurring unitary process or K-Anonymization.

Workflow Manager: interface for managing unit process sequencing workflows.

Fake Datasets Generator: generation of plausible test data in terms of consistency and distribution : sensitive personal data, banking, health ...

Regex Generator: regex test, regex write, automatic regex extraction from sample dataset.

Regex Database: management of a database of custom expressions.

Config Files Generator: generation of configuration files for unitary process.

Sample data manager: test samples and associated configuration files.

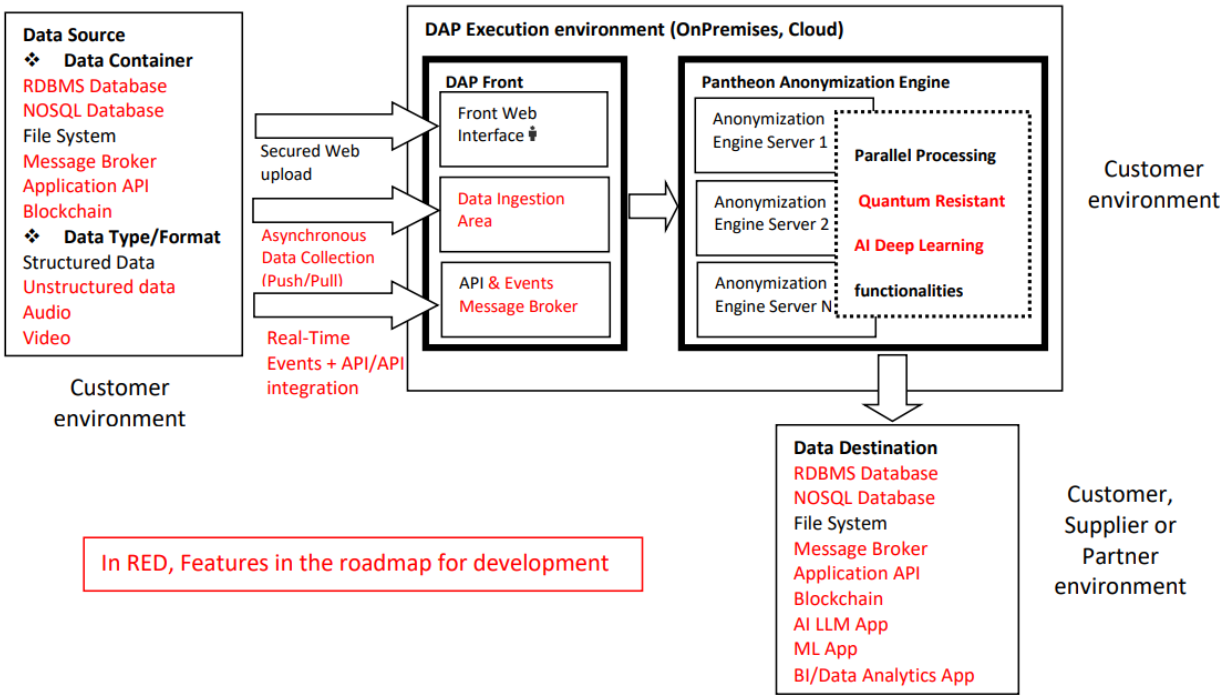
- Effectiveness:

Tokenization is an effective technique for anonymizing sensitive data, providing data security while preserving data utility. However, its effectiveness depends on proper implementation, consideration of contextual information, adherence to regulatory requirements, and integration within the overall system architecture.

That's why DAP allow you to choose the technique you want to apply, among the ones it proposes. We are at your disposal to guide you on the best possible technique, proposed by DAP, during an audit in your company, in order to most effectively anonymize your data flow and then allow you to make the best use of your internal data processing.

3. Data Processing Workflow:

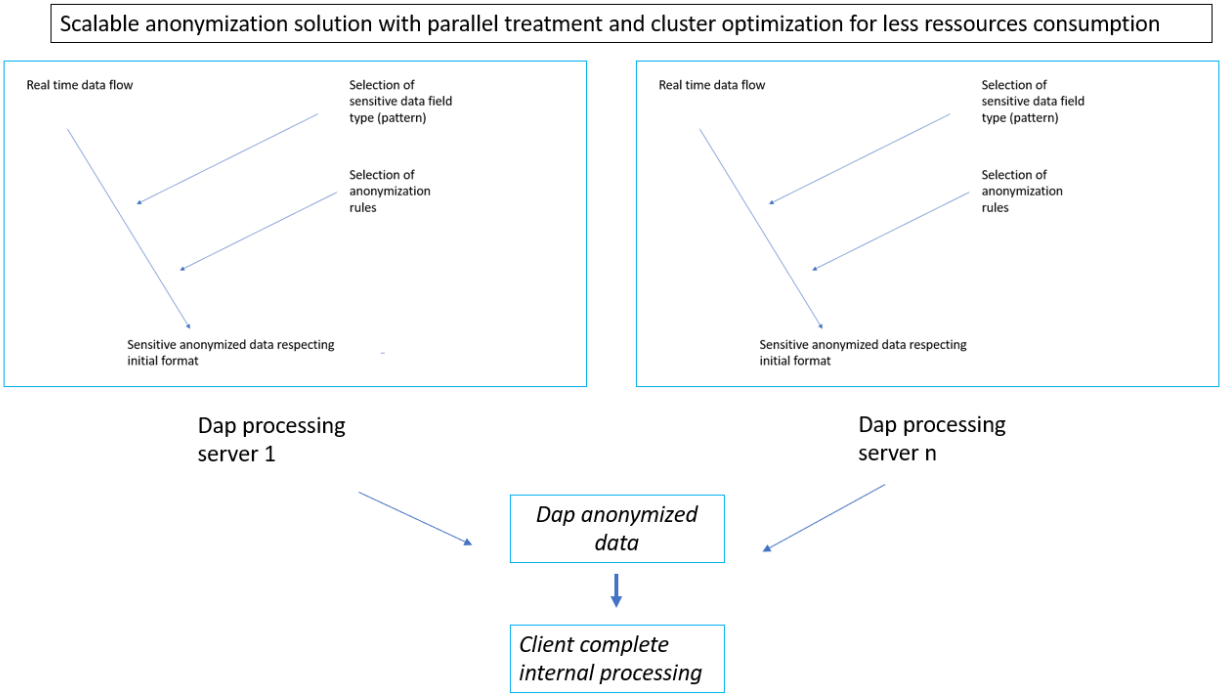
    - Data Flow Diagrams:



DAP software : technical summary

Functional vision:

A scalable anonymization solution with parallel treatment and cluster optimization aims to efficiently process and protect sensitive data by distributing the anonymization workload across multiple parallel processing units within a cluster, ensuring efficient resource utilization and maintaining data privacy.



- Data Transformation:

One example: health data

Data transmitted as a file in .json format.
For each patient, the list of operations performed.
Sensitive personal data:
Forename
Name Social Security Number
Surgeon in charge of the operation

DAP software : technical summary

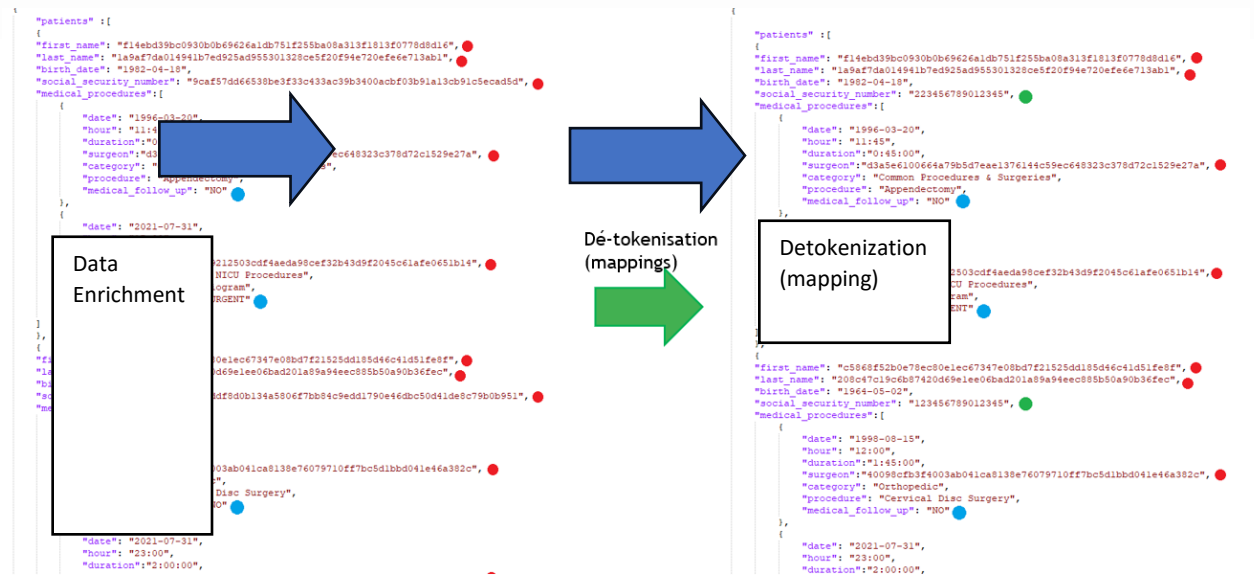

```
{
    "patients" :[
    {
    "first_name": "Sarah",
    "last_name": "Fisher",
    "birth_date": "1982-04-18",
    "social_security_number": "223456789012345",
    "medical_procedures":[
        {
            "date": "1996-03-20",
            "hour": "11:45",
            "duration":"0:45:00",
            "surgeon":"Dr.John Do",
            "category": "Common Procedures & Surgeries",
            "procedure": "Appendectomy"
        },
        {
            "date": "2021-07-31",
            "hour": "17:00",
            "duration":"0:25:00",
            "surgeon":"Dr.Katherine Woods",
            "category":"Neonatal / NICU Procedures",
            "procedure": "Echocardiogram"
        },
    ]
    },
    {
    "first_name": "William",
    "last_name": "Green",
    "birth_date": "1964-05-02",
    "social_security_number": "123456789012345",
    "medical_procedures":[
        {
            "date": "1998-08-15",
            "hour": "12:00",
            "duration":"1:45:00",
            "surgeon":"Dr.Terry Wagner",
            "category": "Orthopedic",
            "procedure": "Cervical Disc Surgery"
        },
        {
            "date": "2021-07-31",
            "hour": "23:00",
            "duration":"2:00:00",
            "surgeon":"Dr.Rebecca Fisher",
            "category":"Cardiac / Cardiothoracic",
            "procedure": "Ablation"
        },
    ]
    },
```


By choice :
Tokenization will only take place on the following sensitive personal data: first_name, last_name, social_security_number.
De-Tokenization will only take place on the following sensitive personal data: social_security_number.

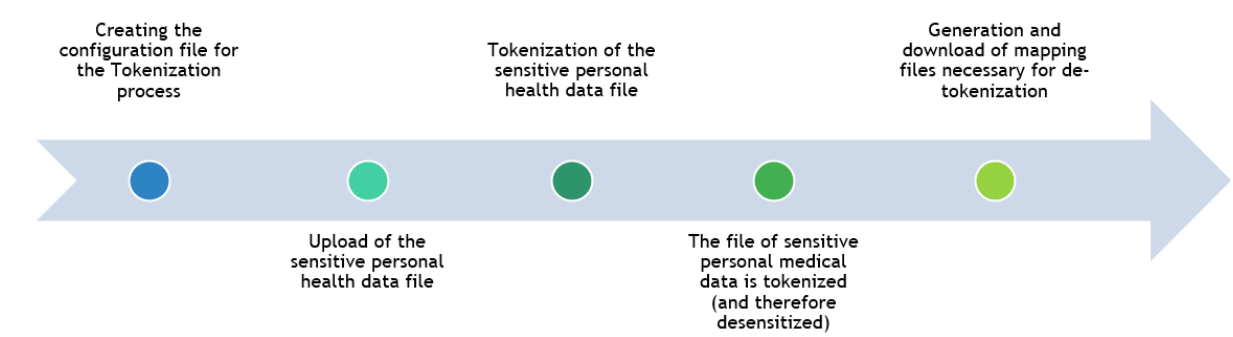
Summary of Tokenization Transformations

DAP software : technical summary

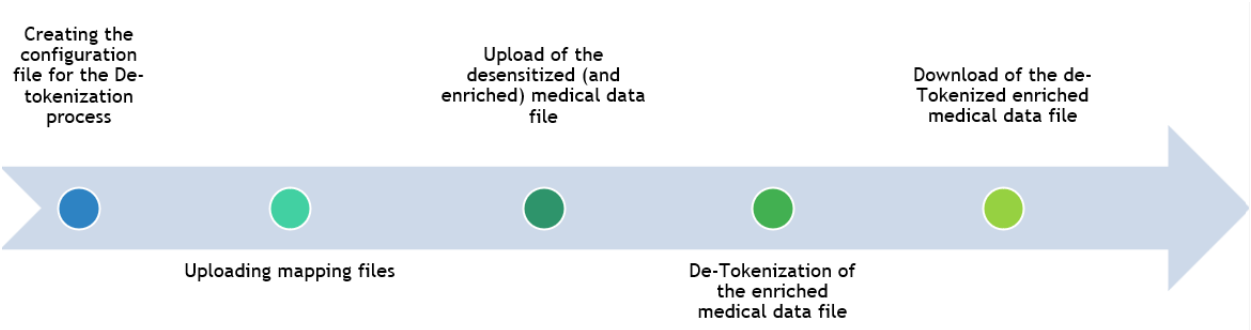
De-Tokenization



Tokenization



De-Tokenization



4. Privacy and Security Measures:

    - Privacy Considerations:

Here are some key considerations about DAP software:

- Data Minimization: collect and process only the minimum amount of data necessary for its intended purpose. It avoids storing or retaining unnecessary personal information.
- Anonymization Techniques: DAP employs robust anonymization techniques to protect personal data. Common techniques include generalization, suppression, pseudonymization, and aggregation. These techniques are applied in a manner that ensures the anonymization can be irreversible and prevents re-identification if this is the technology chosen by the client.
- Privacy by Design: DAP incorporates privacy principles from the initial design stages. Privacy considerations are woven into the software's architecture, data processing workflows, and feature development. This includes implementing privacy-enhancing technologies and adopting privacy-centric defaults.
- Data Security: DAP prioritizes data security to protect the anonymized data from unauthorized access, breaches, and accidental disclosures. It implements appropriate security measures such as encryption, access controls, and secure data storage practices.
- Consent Management: If the process requires user consent for data processing, DAP provides mechanisms to obtain, manage, and document user consent appropriately. Users have clear visibility into the data being collected and processed, and they have the ability to revoke consent if desired.
- Data Subject Rights: DAP facilitates the exercise of data subject rights, such as the right to access, rectify, or erase personal data. It provides mechanisms for individuals to submit data subject requests and enable the software's users to respond to these requests in a timely and compliant manner.
- Auditing and Logging: DAP is planned to maintain in the near future comprehensive audit logs and logs of data processing activities. This helps in ensuring accountability, monitoring system activities, and providing an audit trail for compliance purposes.
- Regular Assessments and Reviews: DAP undergoes periodic assessments and reviews to identify and address any privacy or compliance gaps. This involves conducting privacy impact assessments (PIAs) or engaging external auditors to evaluate the software's privacy controls and practices.
- User Education and Awareness: The software provides a web learning platform, adequate user documentation, training materials, and support to help users understand the privacy features, settings, and best practices for data protection. This includes educating users on the importance of anonymization and responsible data handling.

DAP software : technical summary

- Security Measures:

Pre-encryption : Data is encrypted with the AES-GCM algorithm, as used in US military-grade encryption, before the transfer to ensure their integrity and confidentiality, then deposited in a digital safe repository dedicated to each customer account using a multi-tenant approach.

Encryption : DAP as the unique particularity to let the user choose the type of algorithm he wants to use, selected in the proposed list during the anonymization processing. The main tokenization algorithms proposed by DAP are SHA256 and SHA 512 (US federal government processing standard for civil use) and AES-256 (military grade encryption standards).

Access control:

Efficient access control is crucial for big data anonymization software to ensure that only authorized users have access to sensitive data and the anonymization functionalities. Here are some components that can contribute to an effective access control mechanism:

- User Authentication: Dateligens team has implemented a robust user authentication system to verify the identity of users accessing the software (username/password authentication). A multi-factor authentication (MFA) is in development. The design of DAP allows too integration with existing authentication systems like LDAP or Active Directory.
- Audit Logging and Monitoring: DAP allows a comprehensive logging and monitoring functionalities to track user activities, access attempts, and system operations. This helps in identifying any unauthorized access attempts, detecting anomalies, and providing an audit trail for compliance purposes. Real-time alerts can also be set up to notify administrators of any suspicious activities. (RoadMap 2024/2025)
- Encryption and Secure Communication: Dap is a dynamic anonymization solution that ensures that data is encrypted both at rest and in transit. This includes encrypting sensitive data stored in databases or file systems and using secure communication protocols (such as HTTPS) for data transmission between different components of the software.

Logs and audit logs (in development, RoadMap 2024/2025)

- User Activity: DAP records user actions and operations, including login/logout events, data access attempts, anonymization configuration changes, and any other relevant user interactions within the platform. This helps establish accountability and traceability of user activities.
- Data Access and Usage: DAP log details about data access, such as the dataset accessed, specific columns or fields accessed, and the purpose of access. This information helps monitor and track data usage, ensuring that only authorized individuals or processes access the data.
- Anonymization Processes: DAP capture details of anonymization processes performed on the data, including the techniques used, parameters configured, and the specific data elements or columns anonymized. This information is crucial for auditability and validating the effectiveness of the anonymization techniques applied.

5. Technical Specifications:

    - Supported Data Types: Actually, files in AVRO format (Big Data pivot format), encrypted Json , Json, Csv, Xml, or plain text, are supported. In the coming months we will also cover the Microsoft office formats (word , xls , ppt ...) , Acrobat PDF, Apache OpenOffice as well as unstructured data and multimedia.
    - Performance: Unlike traditional static anonymization methods, DAP utilizes advanced algorithms and machine learning techniques to dynamically anonymize data **in real-time**. By adapting the anonymization techniques based on the context and **sensitivity/ granularity** of the data, DAP ensures optimal privacy protection while preserving the usefulness of the data for analytics.
    - Scalability: The design and encoding choices, using SPARK and SCALA ensures a very high level of native scalability. To develop?
    - Metrics:

PoC SG : 100 Gbytes/hour on Apache logs. Encrypted data on the cluster with Ranger TDE EZ (cold and in-transit data).

AWS : 250 Gbytes/hour. To add more details about the chosen AWS environment

With AMD x10 accelerator board, on AWS theoretically at 2.5 Tbytes/hour.

6. Integration and Interoperability:

    - Compatibility: Linux OS, possibility of using in any environment with virtual machine including Linux
    - APIs and Interfaces: in development (particularly, SAP connector).

7. Quality Assurance and Testing:

    - Testing Procedures:

Throughout the testing process, Dateligens maintain detailed test documentation, record test results, and track any identified defects or issues. Collaboration between with the members of development team helps to resolve efficiently any identified problems and to conduct retesting as needed.

We adapt the testing approach based on the specific requirements, use cases, and deployment mode (SaaS or on-premise) of DAP. The testing procedure outlined above serves as a general framework to ensure thorough testing and validation of DAP's functionalities, performance, security, and compliance.

    - Quality Control:

Quality control for a SaaS or on-premise big data anonymization software like DAP involves a range of activities to ensure its reliability, functionality, performance, security, and compliance.

It begins with thorough requirements management, followed by comprehensive test planning and execution to validate the software's functionalities and performance. Bug tracking and defect management processes are employed to capture and address issues identified during testing or reported by users. Configuration management and release management practices ensure proper version control and smooth deployment of new software versions or updates.

Continuous monitoring of performance, security audits, and compliance checks help maintain a high level of software quality. User feedback and satisfaction are valued, and efforts are made to provide comprehensive documentation and training resources. The aim is to foster a culture of continuous improvement and deliver a reliable and effective anonymization solution for big data.

8. Regulatory Compliance:
    - Compliance Statements: DAP adhere totally to relevant data protection regulations and industry-specific standards. This includes compliance with regulations such as the General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), or other applicable data protection laws.
    - Certifications: DAP has been developed in accordance with the principles of the ISO 27001 standard
9. User Documentation:
    - User Manual: A multilingual user manual is available online.
    - Troubleshooting and Support: A 24/7 support is available through the maintenance option, by the intermediary of a third part support team (international integrator and support services company).