General Intelligence and
Security Service
Ministry of the Interior and
Kingdom Relations

Deployment Advisory BSPA Assessment of Virtual Data Lake

Date February 29, 2024

Colophon

Our reference number	:	9a8e20a9-or1-1.0
	:	T +31 (0)79 320 50 50
	:	F +31 (0)70 320 07 33
	:	P.O. Box 20010 2500 EA The Hague The Netherlands

Copy number	:	
Author(s)	:	NLNCSA
Number of enclosures	:	0

1 Table of contents

2 Statement of Conformity

Hereby is stated that evaluation has demonstrated that the product:

Virtual Data Lake
from
Roseman Labs

is in conformity to:
BSPA Security Evaluation Target for Virtual Data Lake, Version 1.1, 28.02.2024

as demonstrated by:
Secura located in Amsterdam, Netherlands

Applying:

NL Scheme for Baseline Product Assessment, 14.09.2017

This Statement of Conformity (SoC) is part of the Deployment Advisory (DA) and is only valid if the recommendations and obligations in the DA are being followed.

The DA determines the conformity of the product with its SET and the effectiveness of the security features offered by the product.

This SoC relates only to a specific version of a product. If the product is changed, this SoC is not applicable anymore. Newer versions of the products need to undergo the BSPA process anew to obtain a new Statement of Conformity.

Issuance of a SoC is no guarantee that the product is free from security vulnerabilities. Also, a SoC is not an endorsement of the IT product by NLNCSA and no warranty of the IT product by NLNCSA or by any other organisation, is either expressed or implied.

Issue date: 19e March 2024

B. Dunnebier

Head of the National Communications Security Agency of the General Intelligence and Security Service

3 Introduction

3.1 Scope

Product name	Virtual Data Lake
Product version	• Portal: v2.2.0 • Cranmera: v1.8.0 • Crandas: v1.8.0 • VDL: v1.8.0
Product category	04 – Media and file security
Evaluation criteria and version	BSPA_D_01_NL_Scheme_for_Baseline_Product_Assessment 14-09-2017.
Vendor	Roseman Labs
Overseer	General Intelligence and Security Service Ministry of the Interior and Kingdom Relations Netherlands National Communications Security Agency (NLNCSA) P.O. Box 20010 2500 EA The Hague The Netherlands
Evaluation facility	Secura B.V. Vestdijk 59, 5611 CA EINDHOVEN The Netherlands Herikerbergweg 15 (Apollo Building, 3rd floor), 1101 CN AMSTERDAM, The Netherlands

The goal of this Deployment Advisory is to inform consumers on:

the specific use case for which the product has been tested,
the manner in which the product has been tested and the limitations of this test process,
the level of security provided by the product when used according to the prescribed use case,
the residual risks of the product when used according to the prescribed use case.

This Deployment Advisory also gives guidance to users and/or administrators on how to securely use and configure the product.

3.2 Disclaimer

This advisory and associated Statement of Conformity applies only to the specific version of the product in its evaluated configuration (see section 4.1).

The Statement of Conformity is not a guarantee that the product is free from security vulnerabilities. Neither is it a guarantee that the product protects against adversaries with a high attack potential like (for example) intelligence organisations, organised (cyber-) crime organisations, terrorist organisations and capable security researchers.

Exploitable vulnerabilities may be discovered after issuance of the Statement of Conformity. The organisation or individual using the product should check regularly if security vulnerabilities have been discovered and whether updates are provided by the vendor. Installation of updates must be compliant with the risk management policy and risk appetite of the organisation or individual using the product. Updates should only be installed if there is sufficient trust or assurance that they improve the security of the product.

This advisory is supplementary to the instructions and documentation of the vendor. It is still necessary to consult these before installing and using the product.

All product or company names used in this report are for identification purposes only and may be trademarks of their respective owners.

3.3 Certification scheme

The Dutch *Baseline Security Product Assessment* evaluates the security features of hardware and software security products for use in the "internal but unclassified" domain. The main goals of the evaluation are:

- Establishing that the product conforms to the security specification in the Security Evaluation Target,
- Establishing the effectiveness of the security features offered by the product,
- Establishing that the product has a limited impact on the security of the host system.

A positive Statement of Compliance and Deployment Advisory enables a government organization to make a better substantiated decision if these products are applicable in an ICT-infrastructure at the BBN1 or BBN2[1] level of the Government Baseline for Information Security (Baseline Informatiebeveiliging Overheid, BIO, Versie 1).

BSPA is intended for "internal but unclassified" information where there is no need for the product to be resistant against adversaries with a high attack potential[2], like intelligence organisations, organised (cyber-) crime organisations, terrorist organisations and capable security researchers

Documentation and procedures of the Baseline Security Product Assessment scheme are available at the following Internet site: www.aivd.nl search for "BSPA".

3.4 Copyright information

3.5 Contact information

All correspondence regarding this deployment advisory should be addressed to:

General Intelligence and Security Service
Ministry of the Interior and Kingdom Relations
NLNCSA
P.O. Box 20010
2500 EA The Hague
The Netherlands
bspa@nlncsa.nl

[1] To attain the BIO BBN2 level, a government organization will at least need:
- additional measures for detection of cyber-attacks by adversaries with high attack potential (BIO 12.4.1.3 and BIO 13.1.2.1),
- an additional and deeper assessment of the cryptography of the product (BIO 8.3.1.2 and BIO 13.1.2.3).

[2] BSPA is not designed and intended for the BIO BBN3 level (and higher classifications and threat levels). For the assessment security products for the BIO BBN3 level a national evaluation scheme exists, with different characteristics.

4 Product overview

4.1 Product version

This deployment advisory is limited to the following versions of the Virtual Data Lake:

- Portal: v2.2.0;
- Cranmera: v1.8.0;
- Crandas: v1.8.0;
- VDL: v1.8.0.

4.2 Product identification

The product versions can be identified in the following manner:

- **VDL and Crandas**
  The version of the VDL and Crandas can be retrieved by running the below stated script against the environment:

```
[7]: import crandas as cd

     #version of the VDL:
     print("VDL version:", cd.base.session._get("/version").json())

     #version of crandas
     print("Crandas version:", cd.base.session.version)

     VDL version: {'version': 'v1.8.0'}
     Crandas version: v1.8.0
```

For Cranmera and the portal version, two statements were provided by Roseman Labs:

We use 'cranmera' and 'VDL' interchangeably. Although in the code base 'vdl' is an app of 'cranmera', cranmera is never deployed alone. Therefore, VDL version == cranmera version.

While the portal does not yet show its version number to the end-user (this is on our backlog), we are always very explicit to our customers (either on-prems or SaaS) about what version we are hosting for them or which version they should pull. When we have a new version, we provide a changelog and a timeline of when their SaaS environment will be updated.

4.3 Assumptions

The following assumptions are considered relevant for the environment:

- **AS.Server_nodes_securely_configured.** The server nodes are securely configured.
- **AS.Server_nodes_are_tamper_free.** The server nodes have not been tampered with by node administrators after configuration (including cache files stored locally on disk).
- **AS.Binaries_verified.** The VDL installation binaries have not been tampered with by the participant which used them for setting up their VDL node.
- **AS.Participants_do_not_collude.** Restoring the secrets from the secret shares requires majority of participants/users to work together. It is assumed that this does not happen.
- **AS.Well_executed_segregation_of_duties.** The duties between the participants/users are well-defined and executed. This is important to ensure that shares are not combined by one party, which reveals input data.
- **AS.Script_approvers_well_intentioned.** The script approvers (See section 3.3 of the SET) for the organizations are well intentioned and do not sign queries which do not have a legal basis or are not in the interest of the participants.
- **AS.Administrators_trustworthy.** Administrators are trustworthy towards their own domain.
- **AS.Key_material_secure.** PKI keys, passwords, two factor authentication, and other key material is used and stored in a secure manner with "good housekeeping practices".

4.4 Security functions

The following security functions of the Virtual Data Lake have been verified by a time-boxed and lightweight pentest:

- **SF.Data_confidentiality.** Input data is encrypted and masked locally at each organization. The data is partitioned into parts that do not disclose any information about the input data. These parts are then sent to the different nodes in the virtual data lake. The input data remains confidential during computation. The column names, also known as the metadata, are sent and stored in plaint text. The confidentiality applies to the input data only.
- **SF.Only_approved_scripts_executed.** Only agreed upon queries can be executed. These queries have been signed by script approvers.
- **SF.Authorization_on_web_portal.** Only authorized entities can access the web portal.
- **SF.Query_output_access**. Only authorized entities have access to query output. This can be a subgroup of the participants.
- **SF.Data_in_transit_encryption.** The VDL nodes are pairwise interconnected via TCP/IP transport layer security v.1.3.
- **SF.Secure_against_minority_of_passively_corrupted_actors.** VDL is secure against a minority of passively corrupted parties. Thus, the VDL binaries on the servers have not been tampered with, and a minority is not able to read the input of other participants.

4.5 Product description

The scope of the evaluation includes the Virtual Data Lake (VDL) by Roseman labs. This is a software solution for performing analysis on combined datasets without sharing the content of the individual datasets between the participating parties. The core of this product is a technique called Secure Multi-Party Computation (MPC) which is carried out by the MPC engine: Cranmera. The VDL is an abstraction level above this engine and can be used to interact with the combined database like any other database. The input data is encrypted locally at each party. After which it is partitioned into parts which do not disclose any information about the input data, and distributed over the MPC servers in the VDL. These parts are called "secret shares" and reveal no information about the data individually. This technique is based on the Ben-or, Goldwasser, Widgerson (BGW) protocol with Shamir's Secret Sharing algorithm.

A collaboration based on the VDL is governed by a group of trustees that agree upon the queries that can be executed, output that can be generated and the individuals that can access the VDL. This is to ensure that no data is being extracted beyond the scope of the collaboration.

Figure 1, which is stated below, displays an example of the VDL setup:

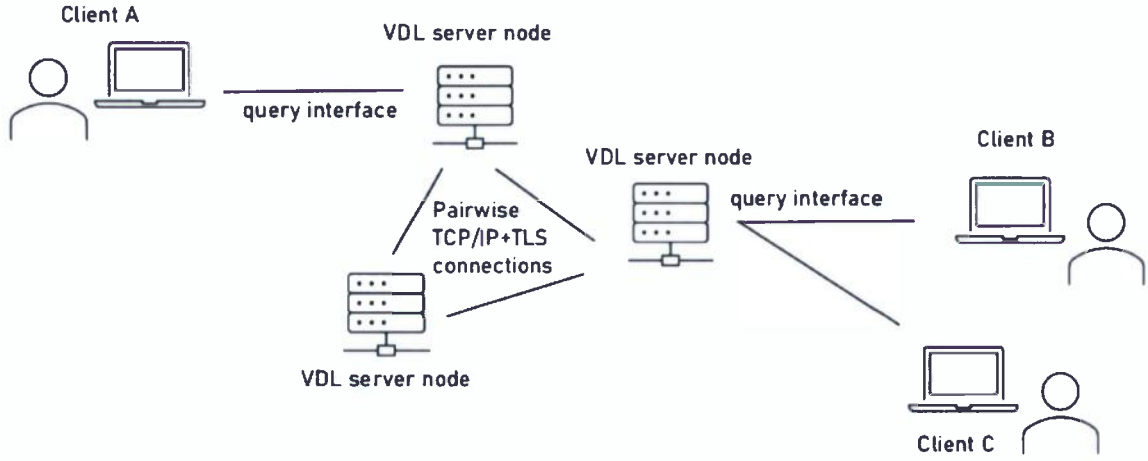

Figure 1. Example VDL Network

Queries can be executed on the VDL after the datasets have been combined virtually using the BGW based protocol with Shamir secret sharing. This splits the dataset into secret shares which are encrypted with the public key from the node on which it will reside. This encryption takes place before the data enters the VDL. Therefore, shares distributed by an input node are protected from the input node.

A script approver signs a query with their private key to approve it for usage on the VDL. Then, a data analyst can connect to the VDL either via the web portal or by using the Crandas python package and execute the query. Each node holds the public keys for script approvers and can therefore check every script for approval by all assigned script approvers before execution. There is a consensus mechanism that enforces that all VDL server nodes must agree in order to collaboratively execute the query, otherwise, the query will not be executed.

The list of script approvers is established at the initialization of the VDL network and is configured via a registration procedure. This registration process is a three-step process and is performed via a web-based user interface, which is referred to as the portal.

4.5.1 Registration process

Step 1: Deployment
A signing key-pair is generated and the portal holds the private signing key. The VDL server nodes are provisioned and configured with the public verification key via a persistent configuration file. Thus, at this stage of the process, the portal has authorized access to the VDL servers.

Step 2: Registering the primary administrator
The primary administrator generates a key pair client-side in the browser. The private key is neither stored on, nor communicated to the server. The portal signs the primary administrator's public key and registers it at the VDL server nodes by sending a configuration command to the VDL server nodes. At this stage, all servers authorize commands against the specified set of public keys, and the portal no longer has authorized access to the VDL servers. This also implies that the output of executions are only accessible to approved participants.

Step 3: Registering the approvers
The script approvers generate their key pairs client-side in the browser and the primary admin transfers its approval power to the final set of script approvers by means of the same configuration command as used in step 2.

4.5.2 Deployment options

A VDL can be deployed in three different ways:

- Centralized: All servers are deployed in a single organization and segregation of duties is implemented between administrators within the same organization.
- Distributed: All servers are deployed in logically and legally separated environments. Thereby, there is clear segregation of duty since the administrators are part of different organizations. A distributed deployment can exist of a combination of cloud and on premise hosted servers.
- SaaS: All servers are deployed in a SaaS environment that is managed by Roseman Labs. Here, segregation of duties is implemented between Roseman Labs administrators and servers are hosted in separate EU cloud providers.

4.5.3 VDL authorization mode

The VDL can either be in authorized or unauthorized mode. When in unauthorized mode, queries can be run without approval on the VDL nodes. This mode is meant for interactive exploration and is therefore intended for environments which only contain dummy data. For production environments, the authorized mode is recommended for VDL. In this mode, all queries have to have a prior approval by a fixed set of script approvers before it can be executed.

4.6 Product category

The product belongs to the following pilot BSPA category:
04 – Media and file security

4.7 Product configuration

As mentioned within section 4.5.2. Deployment options, the VDL can be deployed in three different solutions. This BSPA was tested from a SaaS perspective.

No configuration options were altered unless it was necessary to test one of the security features in scope of the assessment.

5 Tested security features for the product

The security features that are part of the BSPA for the Virtual Data Lake are listed in Table 1.

Security features part of Virtual Data Lake	Security feature part of BSPA scope.	Additional information
SF.Data_confidentiality	Yes	-
SF.Only_approved_scripts_executed	Yes	-
SF.Authorization_on_web_portal	Yes	-
SF.Query_output_access	Yes	-
SF.Data_in_transit_encryption	Yes	-
SF.Secure_against_minority_of_passively_corrupted_actors	Yes	-
The following features triggered by or related to the security feature in scope that might be expected by a user, are explicitly not included in the scope of the evaluation:		
The strength of used encryption algorithms	No	The actual implementation of the encryption algorithms is out of scope for the evaluation. However, a security analysis of the cryptographic building blocks is available for interested stakeholders (under NDA), and a public vulnerability analysis is applicable.
The security of TLS connection	No	The actual implementation of the TLS protocol is out of scope for the evaluation, however a public vulnerability analysis is applicable.

Table 1. Virtual Data Lake product security features that have been part of this BSPA.

6 Scope and limitations of the evaluation

6.1 Evaluation facility

The security test has been performed by
Secura B.V.
Vestdijk 59,
5611 CA EINDHOVEN
The Netherlands

Herikerbergweg 15 (Apollo Building, 3rd floor),
1101 CN AMSTERDAM,
The Netherlands

https://www.secura.com

6.2 Test duration and time used

The evaluation process has been performed in the period of 24 October 2023 to 8 December 2023.

6.3 Evaluation process and scope

The following time-boxed evaluation process was employed:

- Security Evaluation Target analysis.
- Product installation and deployment analysis.
- Compliance analysis by documentation review.
- Compliance analysis by product testing.
- Vulnerability analysis.
- Usability analysis.

Note: Roseman Labs provided a statement that they are addressing the reported issues that were discovered throughout the BSPA. However, as certain mitigations actions will change the version of the component related to the Virtual Data Lake, this falls out-of-scope of the current BSPA.

6.4 Product procurement, installation and configuration for evaluation

The VDL environment is available by contacting Roseman Labs directly.
For this BSPA, the environment was provided by Roseman Labs as a SaaS solution. Therefore, the product deployment and installation was fully performed by Roseman Labs. The only steps executed by Secura, in order to make use of the VDL environment, were setting-up a direct connection to the environment. The steps for doing so, were provided by Roseman Labs.

For the SaaS environment, after installation of the client, the application is ready for use.
No additional configuration is required.

During the first register of a user, a primary admin account, which is the first (human) administrator which is registered to the system, is created. The primary admin is at this phase of the configuration procedure the sole query approver for the environment, as the set of query-approvers is not yet known when the system is deployed. After initial installation of the product the primary admin can start the registration procedure, which is assigning specific users with approver rights. Note that during the phase where the primary admin is the sole query approver, the system should not yet be used for data uploading/processing tasks.

Once users are assigned with approver rights by the primary admin, the primary admin account loses their own approver rights, which is also known as the bootstrap phase. Each VDL server node thereafter holds a copy of the set of approvers, where each approver is represented in this set by means of its (public) signature-verification key.

7 Instructions and recommendations for users

The following roles and responsibilities are present in a VDL setup:

- Participants: Organizations participating in the data collaboration.
- Input party: A participant that provides data input to the computation.
- Compute party: A participant that runs and administers a VDL node (server).
- Output party: A participant that is eligible to receive computational results.
- Client: A natural person, affiliated to one of the participants that runs queries to upload data, perform data analyses and/or download results, typically from a Python environment.
- Script approver: A natural person which represents the participants that is given the authority to authorize queries and should only do so if running the particular query would be in the interest of the participants, and has a legal basis (e.g. under the GDPR). The script approver can authorize a query by digitally signing it.

It is important to note that every participant could fulfil multiple roles simultaneously. Therefore, the recommendations for administrators and users are combined into one chapter.

7.1 Procurement of the product

No additional procurement risks for the user were identified.

7.2 Installation of the product

No additional installation risks for the user were identified.

7.3 Configuration of the product

Two configuration risks were identified throughout the assessment.
The following recommendations were reported:

- Users should perform a coherence check to verify whether the admin page is not publicly accessible. This ensures that the attack surface of the VDL is limited.
- If the user is responsible for the configuration of the web portal they should ensure that version disclosure is not in place within server responses. Version disclosure could assist an attacker to further facilitate attacks.

7.4 Security risks and countermeasures

Recommendations for secure usage:

- Users should set long and complex passwords to ensure that an attacker is less likely to guess valid credentials. Note that Roseman Labs had based their password length and complexity requirements on section 9.4.3.1 of the government's BIO (version 1) document, which states that passwords should be at least eight characters long if no two-factor authentication is being used.
- Users should not disable Multi-Factor Authentication, as it may allow an attacker to compromise user accounts with more ease.
- Recorded scripts should be fully understood by the approver before being approved, to ensure that no ill-intentioned scripts are being approved.

7.4.1 Publicly known vulnerabilities

No publicly known vulnerabilities were identified for the Virtual Data Lake during the conducted publicly known vulnerability analysis evaluation activity.

7.4.2 Independent vulnerability analysis

An independent vulnerability analysis on the components of the Virtual Data Lake was conducted during the evaluation. The following activities were in the scope of the investigation:

- **Attack scenario 1: Authorization bypass**
  To verify whether security issues could be identified for the authorization of users.
- **Attack scenario 2: Unauthorized Roles change**
  To test whether the API implement strict authorization checks in case of critical functionality, such as the alteration of user roles, to verify whether privilege escalation attacks were possible.
- **Attack scenario 3: Public Key Modification**
  To test whether users could modify their own public key after generation, or modify other user(s) public key(s), to conduct impersonation attacks.
- **Attack scenario 4: Unauthorized analysis approval and execution**
  The verify whether the application allowed the execution of unapproved analysis, and whether a user without an Approver role could approve analysis.
- **Attack scenario 5: Account brute-forcing**
  To test whether user credentials and the OTP were susceptible to brute-forcing attacks.
- **Attack scenario 6: Bootstrap process misuse**
  To verify whether the bootstrap process removed the primary admin role after the approvers are set.
- **Attack scenario 7: Server process**
  To inspect the Deamon VDL process.
- **Attack scenario 8: Memory Corruption Analysis**
  To inspect the security mitigations included in the nodes server's binary.
- **Attack scenario 9: Malicious input script**
  To test whether malicious input scripts could be run.

The results of one of the performed tests led to the conclusion that an attacker who is able to discover a memory corruption vulnerability is more likely to exploit this vulnerability due to lacking security mitigations within the binary.

However, this threat scenario does not demonstrate that the product is vulnerable to memory corruption attacks. Therefore, no issues that would directly impact one of the security features of the product were identified throughout the assessment.

7.5 Summary of recommendations

The following recommendations on actions are given to the users of the Virtual Data Lake:

- Users should set long and complex passwords.
- Users should not disable Multi-Factor Authentication.
- Users should perform a coherence check to verify whether the admin page is not publicly accessible.
- Recorded scripts should be fully understood by the approver before being approved.
- If the user is responsible for the configuration of the web portal, they should ensure that version disclosure is not in place within server responses.