**SNYPR Datalake**

**Data source Onboarding Documentation**

**Contents:**

1. **Onboarding request information**
   1. **Initial Information**
   2. **How to use the form**
   3. **Basic log events required for ingestion**
2. **Logging levels for data sources**
3. **Methods of integration**
   1. **Syslog**
   2. **API**
   3. **Database**
4. **Additional Questions**

**1. (a) Onboarding request information – Initial information**

**Purpose:** This form collects information required by the Cybersecurity Engineering team to prepare for the ingestion of data source logs into Data Lake.

**Instructions:** Complete the form and email to: ***CyberSecurityEngineering@AmerisourceBergen.com***

|  |  |
| --- | --- |
| Understanding how a data source is designed and behaves will assist the Cybersecurity Engineering team in determining the best way to ingest these logs into Data Lake. | |
| **Data Source Name:** | **Business Unit:** |
| **Point of Contact:** | **Technical Contact:**       *(If different than Point of Contact)* |
| **Data Source Overview** –*Identify your data source’s general function to provide context to the Cybersecurity Engineering team.* | |
| **Data Source Location** –*Identify where your data source resides. On premise? If so, which data center or physical location? If it is in the cloud, identify the service (ex. Azure, AWS).* | |
| **Data Source Connectivity -** *If not on-premise at an ABC location, does the data source have VPN connectivity back to the ABC corporate environment?* | |
| **Do you have any specific attributes relating to the application data that you think needs to be monitored more?** | |
| **Is logging currently enabled for this data source? If so, please describe how logging takes place and whether this logging is monitored.** | |
|  | |
| **Pick all the connectivity options this data source supports from the following :**    **[Sys]** | |
| **[Syslog – SFTP/SCP – Database Pull – REST API – Others (please specify)]** | |
|  | |
| **Sensitive Data** —*Identify any sensitive data that is processed or stored by the application.* | |
| *Is PHI/PII data processed OR stored by this application?*  Yes  No | **Data Classification:** Select One |
| *If yes, please identify where processed/stored:* | |
| *If yes, has a Risk Assessment been conducted for this data source/application?*  Yes  No | |
| *If yes, do you have target dates for data ingestion that must be met as part of a compliance requirement?* | |
| **Data Type***: identify what type of data will be ingested (web data, SQL, etc.)* | |
| **Additional Information:** | |

**(b) How to Use the Form**

* **Data Source Name –** Input the name of the application/data source
* **Business Unit –** Type the Business Unit in which this data source is under
* **Point of Contact** – Name of the data source owner that will interface with Cybersecurity Engineering
* **Technical Contact** – The technical point of contact for the data source that will interface with Cybersecurity Engineering
* **Data Source Overview -** Identify your data source’s end-to end process flow and what your application does. Optionally, data source owner can schedule a demo of the application or provide architectural diagram illustrating the workflow of the application
* **Data Source Location** – Identify location of data source, whether on-premise, off-premise or in the cloud. If in the cloud, identify which cloud service
* **Data Source Connectivity** – Identify if data source has connectivity to ABC corporate network
* **Logging Enabled** – Identify whether logging is currently enabled for this data source. If so, please explain how logging currently takes place
* **Syslog Forwarding** – Identify whether data source is capable of syslog forwarding
* **API Availability** – Identify whether data source has an API available for incoming connections
* **Sensitive Data –** Identify any PHI/PII data that is processed or stored by the application. Note: If application does not contain PHI/PII data, assign the appropriate data classification type:
  + **Public**: The Public Information level includes AmerisourceBergen information that is in the public domain.
  + **Internal Use**: The Internal Use Only Information level includes AmerisourceBergen information considered to be routine operational information, the release of which must be approved prior to dissemination outside the organization.
  + **Confidential**: The Confidential Information level includes non-public information that either belongs to AmerisourceBergen or belongs to another entity to which AmerisourceBergen owes a duty of confidentiality.
  + **Protected**: "The Protected Information (PI), level includes personally identifiable data in any form (hard copy or electronic) that is subject to state or federal laws and regulations that place restrictions on data use and disclosure."
* **Additional Information –** Additional information that may assist the Cyber Security Team with the Cyber Data Lake onboarding.

**(c) Basic log events required for ingestion**

1. **Operating System(OS) Events**

* start up and shut down of the system
* start up and shut down of a service
* network connection changes or failures
* changes to, or attempts to change, system security settings and controls

1. **OS Audit Records**

* log on attempts (successful or unsuccessful)
* the function(s) performed after logged on (e.g., reading or updating critical file, software installation)
* account changes (e.g., account creation and deletion, account privilege assignment)
* successful/failed use of privileged accounts

1. **Application Account Information**

* successful and failed application authentication attempts
* application account changes (e.g., account creation and deletion, account privilege assignment)
* use of application privileges

1. **Application operations**

* application startup and shutdown
* application failures
* major application configuration changes
  + 1. application transactions:
    2. e-mail servers recording the sender, recipients, subject name, and attachment names for each e-mail
    3. Web servers recording each URL requested and the type of response provided by the server
    4. business applications recording which financial records were accessed by each user

The details logged for each event may vary widely, but at minimum each event should capture

* timestamp
* event, status, and/or error codes
* service/command/application name
* user or system account associated with an event
* Device used (e.g. source and destination IPs, terminal session ID, web browser, etc)
* Message/ Event ID/ Action Performed
* Events related to Data Exchange: File Name, Bytes Received and Bytes sent

**2. Logging levels for data sources**

**Cybersecurity** can suggest recommended log levels for data sources ingesting into the datalake. The table below shows log levels of some of the common data sources integrated into the datalake. If there is a specific application to be ingested which isn’t available in the table below, cybersecurity can provide inputs for recommended levels on request.

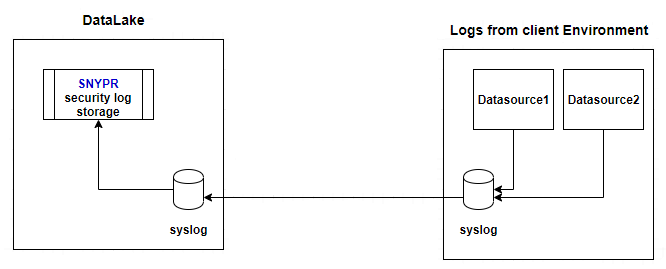
Please be aware that these logging levels are suggested for best case threat management. They are to be in no way mistaken for compliancy. Department specific compliancy for logs should be known and owned by individual departments. Cybersecurity can work with the compliancy requirement to come up with a logging level that can solve a dual purpose of threat management and compliancy.

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Source** | **Available levels** | **Recommended levels** |
|  |  |  |  |
| 1 | AD/DC | Levels : 0 - 5 | 2,3 |
|  |  |  |  |
| 2 | Oracle | Levels : 0 - 6 | 2,3,4,5 |
|  |  |  |  |
| 3 | Apache Web server | Levels : 0 -15 | 4,5,6,7,8,9,10,11 |
|  |  |  |  |
| 4 | Cisco ASA | Levels : 0 - 7 | 4,5,6 |
|  |  |  |  |
| 5 | Ironport | Levels : 0 - 7 | 4,5,6 |
|  |  |  |  |
| 6 | Sonicwall | Levels : 0 - 7 | 4,5,6 |
|  |  |  |  |
| 7 | Barracuda | Levels : 0 - 7 | 4,5,6 |
|  |  |  |  |
| 8 | Palo Alto | Levels : 0 - 4 | 2,3,4 |
|  |  |  |  |
| 9 | Netscalar | Levels : 0 - 7 | 4,5,6 |
|  |  |  |  |
| 10 | Infoblox | Levels : all, on, off | on |
|  |  |  |  |
| 11 | Change Auditor | Levels : 0 - 7 | 4,5,6 |
|  |  |  |  |
| 12 | Netscreen | Levels : 0 - 7 | 4,5,6 |
|  |  |  |  |
| 13 | IBM IDS/IPS | Levels : 0 - 11 | 4,5,6,7,8,9 |
|  |  |  |  |
| 14 | Websense | Levels : 0 -3 | 3 |
|  |  |  |  |
| 15 | Netgear | Levels : 0 - 4 | 2,3 |
|  |  |  |  |
|  |  |  |  |
| 16 | Beyond Trust | Levels : 0 - 2 | 1,2 |
|  |  |  |  |
| 17 | McAfee | Levels : 0 - 3 | 1,2 |
|  |  |  |  |
| 18 | MySql | Levels : 0 - 5 | 2,3,4 |
|  |  |  |  |
| 19 | Application server - Apache Tomcat | Levels : 0 - 6 | 3,4,5 |

**3. Methods of Integration**

**a) Syslog**

Syslog is the most common method of ingesting logs into the datalake. If syslog is configured in the client servers, then it is just a matter of getting the ip address and port number of the server to the syslog in the datalake. We can then parse and ingest the logs from syslog into SNYPR which is our log storage and cyber analytic engine.

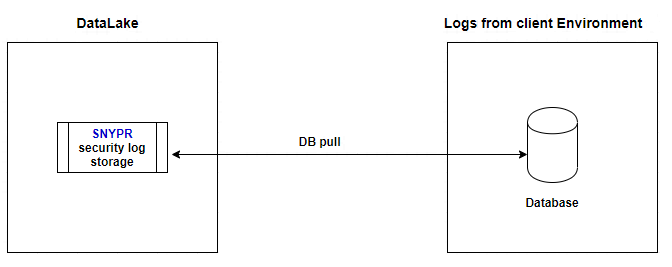


**Fig 1. Log forwarding through syslog**

* Ensure that firewall rules (and VPN connectivity, if applicable) are in place to route traffic to the datalake
* Enable syslog forwarding to the datalake. TCP connections are recommended over UDP to minimize possibility of data loss.

**b) Database**

SNYPR can directly pull data from databases. Database queries from SNYPR can be used to connect to the client database and pull information on a periodic basis. Normally user and HR information are stored in databases. This information will be useful to correlate an event back to a user. Some of the common databases that SNYPR can connect to are MySql, Oracle, Microsoft SQL etc.

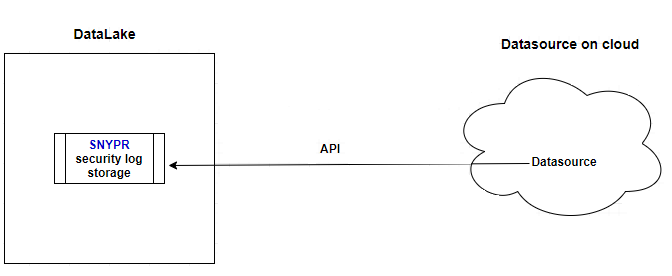


**Fig 2. DB pull for ingesting data from database**

* Provide read-only service account for datalake to connect to the database server.
* Provide information on the table or view where the data is being stored.
* Ensure that firewall rules (and VPN connectivity, if applicable) are in place to route traffic from/to the datalake

**c) API**

SNYPR can make API calls to applications running on cloud. This is another way of ingesting data and particularly used for cloud based applications.

****

**Fig 3. API calls used for data source on cloud**

* Provide information on API endpoints, attributes and example queries with output.
* Ensure that firewall rules (and VPN connectivity, if applicable) are in place to route traffic from/to the datalake.
* Provide authentication information to complete the connecrto the API (service account details, auth tokens, etc.)

**d) SFTP/SCP**

SNYPR can also be set up to receive via SCP or SFTP. This method will permit us to collect the logs periodically, when a more real-time method like syslog is not available. To enable this feed:

* Ensure that firewall rules (and VPN connectivity, if applicable) are in place to route traffic to the datalake
* Enable periodic push of log files via SCP/SFTP through a datalake service account.

**4. Additional Questions**

**(i) Questions Cyber team can help/ assist with**

a) Cybersecurity can give pointers to logging levels for datasources that are not there in the list on section 2 based on the inputs given by the technical team.

b) Cybersecurity can also help with deciding the best course of integration into the datalake (Usually done during the first call).

**(ii) Questions Cyber team cannot help with**

a) Impact of increasing the logging levels on the application.

b) For applications not listed in section 2, application team are resposnsible to analyze and identify available logging levels which cybersecurity can help validate.

c) Installing syslog/ other parameters in the environment.

d) Application team are expected to have local logging enabled before reaching out to cybersecurity for datalake onboarding.