Seqrite Data Privacy Release 2.2 Sizing Guide

Introduction

This guide provides detailed instructions for sizing the deployment stack for Seqrite Data Privacy Release 2.2. The resource requirements for each component in the scan pipeline are influenced by various factors, which are outlined separately for each component.

Factors Affecting Resource Allocation and Scalability

Scaling and sizing for Seqrite Data Privacy depend on several factors:

1. **Number of Data Source Instances**
2. **Type of Data Sources**
3. **Size of Data to Be Scanned**
4. **Number of Endpoints/Devices**
5. **Number of Active Classifiers**
6. **PII Density**

Critical Components

The following components are critical in the scan pipeline and may require additional resources as data size increases. Resource requirements for these components are specified individually:

Sr. No.	Component	Criteria
1	Endpoint Data Discovery Engine	Number of instances depends on the number of endpoints/devices to be supported.
2	Data Discovery Engine	Number of instances depends on the number of data-source instances configured.
3	Data Classification Engine (Executor)	Number of executors depends on the number of Endpoint DDEs and DDEs configured in the system.
4	Data Classification Engine (Driver)	Resources depend on the number of executors configured.
5	Mongo Database (Single Instance)	Resources depend on the amount of data stored and PII density.
6	Message Queue	Resource allocation depends on the number of DDEs.

Assumptions

1. The number of active classifiers is assumed to be 10 or fewer.
2. Low internet bandwidth may increase scan times when fetching data from remote data sources.
3. The Mongo database used as local storage is a replica set with 2 nodes; the resource requirements provided are for a single node.

Availability Schedule Guidelines

The configuration of the availability schedule for each data-source instance may affect the count and resource requirements for Data Discovery Engines. Follow these guidelines:

1. Ensure that scan availability times for data-source instances do not overlap.
2. Each scan engine handles only one data source at a time.
3. Multiple scan engines can be configured in the cluster.
4. The number of scans running concurrently depends on the number of scan engines.

Scaling Factors for Critical Components

The scaling factors for critical components are as follows:

Sr. No.	Component	Criteria
1	Endpoint Data Discovery Engine	One instance per 100 endpoints/devices.
2	Data Discovery Engine	Minimum of 3 instances for up to 8 data-source instances; add one additional instance for every 3 additional data-source instances.
3	Data Classification Engine (Executor)	Minimum of 3 executors; one additional executor per additional Endpoint DDE or DDE.
4	Data Classification Engine (Driver)	Single instance required; resources depend on the number of executors.
5	Mongo Database (Single Node)	Configuration provided supports up to 1 TB of data with up to 2% PII density.

Please note, here Instance typically refers to a single deployment of a component. For example, 1 instance of the Data Discovery Engine means one running deployment of that component, and the same applies to other components like Endpoint Data Discovery Engine, Data Classification Engine (Executor), etc.

Resource Requirements for Critical Components

The resource requirements for each instance of the critical components are as follows:

Sr. No.	Component	CPU (vCPUs)	Memory (GB)	Storage (GB)
1	Endpoint Data Discovery Engine	0.4	2.0	5.0
2	Data Discovery Engine	1.0	4.0	7.0
3	Data Classification Engine (Executor)	4.0	10.0	N/A
4	Data Classification Engine (Driver)	1.0	6.0	N/A
5	Mongo Database (Single Instance)	8.0	20.0	60.0

Resource Requirements for Remaining Components

The resources required for remaining components are:

- **Total CPUs:** 35 vCPUs
- **Total Memory:** 67 GB
- **Total Storage:** 750 GB

Notes

1. Configuration for each customer will be calculated based on specific information provided by the customer.
2. Deployment scripts will be modified according to the configuration before use.
3. Total CPUs and memory required can be provided as a set of virtual machines or physical machines.