



Data Security Automation for Data Warehouses

Reference Architecture

Table of Contents

Table of Contents	3
Executive Summary	4
Audience	5
Purpose	5
Redshift and Snowflake Overview	6
Snowflake	6
Amazon Redshift	6
LightBeam Data Privacy Automation Platform	7
Main Dashboard	7
Attribute View	8
Entity View	9
LightBeam Operational Phases	9
Detect	10
Enforce	11
Policies	11
Permissions	11
Alerts	12
Redaction	12
Data Privacy Automation for Data Warehouses	13
Automate	13
Connecting Data Warehouses to LightBeam	14
Connecting Amazon Redshift to LightBeam	14
Step 1: Basic Information	15
Step 2: Connection Details	16
Step 3: Scan Settings	17
Connecting Snowflake to LightBeam	18
Step 1: Basic Information	18
Step 2: Connection Details	19
Step 3: Scan Settings	20
Conclusion	21
Appendix	21
Revision History	21
About LightBeam	21

Executive Summary

The key in meeting the requirements of today's data security and privacy regulations, and protecting sensitive information (PII/PCI/PHI) from unauthorized use and disclosure lies in understanding and managing the use of sensitive information within an organization's data environment. Spread across a multitude of repositories and application data sets, sensitive information use can be difficult to manage through written policy alone.

We at LightBeam.ai believe the best way to implement policy across an organization is to supplement written policy with procedures for technical controls designed for specific applications and functions. By working with our clients we have developed several application and function specific controls focusing on discovering, analyzing, and enforcing control over the use of sensitive information within popular applications.

Data Warehouse

With the advent of cloud computing and increasing data analytics needs, the concept of large scale databases has given birth to new data platforms known as data warehouses. A data warehouse is optimized for querying and analysis of structured data that has already been processed, cleaned, and transformed for specific business purposes. It integrates data from multiple sources into a unified schema for easier querying and reporting. Data warehouses are used for generating structured reports, and dashboards, and conducting business intelligence operations. Data warehouses store data in a structured format with predefined schemas. Data is transformed, cleaned, and organized into a schema that is optimized for analytics and querying (schema-on-write). Data warehouses are optimized for online analytical processing (OLAP) and typically use SQL queries for data retrieval and analysis. They are less flexible in handling raw data and more focused on providing fast query performance on structured data. Business analysts, executives, and operational users who need structured and consistent data for reporting, dashboards, and decision-making primarily use data warehouses. Applications include business intelligence, performance reporting, and operational analytics. Two very popular data warehouses are Amazon's RedShift and the Snowflake platform.

LightBeam's AI-driven platform engine, Spectra, can easily be configured for attribute scanning in these platforms to automatically discover, analyze, and enforce compliance policies regarding the use of sensitive information. By finding and either raising alerts, overwriting, or deleting data that inappropriately contains PII/PCI/PHI, organizations can reduce data security risk and meet retention requirements for data that is no longer needed. By then automating the execution of these control policies, Data Protection Officers can develop custom rule sets that continually scan, monitor, and control how sensitive information is used and controlled within Redshift and Snowflake repositories. The details of how this happens are discussed below.

Audience

This document is intended for organizations that have implemented either Redshift or Snowflake and whose information processing uses personal information. It is meant for both technical and non-technical audiences. CISOs, Information Security teams, Privacy Officers, and Support leaders within organizations overseeing the use of PII/PCI/PHI in data warehouses will find this reference architecture useful in automating data security and privacy controls.

Purpose

This document provides greater details on the problem of storing, processing, and accessing sensitive information within data warehouses and how LightBeam can be used to manage the use of PII/PCI/PHI and reduce the risk posed by data duplications or inappropriate and long term storage of sensitive information. Although this document is primarily about Redshift and Snowflake it applies in principle to other more traditional data warehouses and databases.

Redshift and Snowflake Overview

Both Amazon Redshift and Snowflake are considered data warehouses with these characteristics.

- Data warehouses like Amazon Redshift and Snowflake are optimized for storing and querying structured data.
- They typically enforce a predefined schema and structure for data storage, which is optimized for analytical querying and reporting.
- Data is cleaned, transformed, and organized into a schema that facilitates efficient data retrieval and analysis.
- These platforms are designed to handle large volumes of structured data and provide fast query performance for business intelligence and analytics applications.

Snowflake

- Snowflake is a cloud-based data platform that offers data storage, processing, and analytics solutions.
- Snowflake is designed to help organizations manage and analyze large amounts of structured and semi-structured data.
- Snowflake's features include a SQL query engine, Cloud architecture and is designed for the cloud that can work on any of the three major clouds.

Amazon Redshift

- Amazon Redshift is a fully managed data warehouse service provided by Amazon Web Services (AWS).
- It uses columnar storage and parallel query execution to handle large-scale data analytics workloads.
- Redshift is best suited for online analytical processing (OLAP) workloads where structured data is analyzed for business insights and reporting.

Data stored in data warehouses can contain sensitive information that is needed to complete transactions, process orders, or complete other activities. Used across a multitude of business types, data warehouses can support many data and process types. Many customer interactions require the use of SI (sensitive information) and the type of SI will vary by process type. While processing a transaction, SI may be required, however, long term storage of SI in old tables and databases creates risk and should be

addressed. Guided by the data protection principle to only keep SI for as long as is needed, LightBeam.ai has developed controls specific for how Redshift and Snowflake uses and retains personal information.

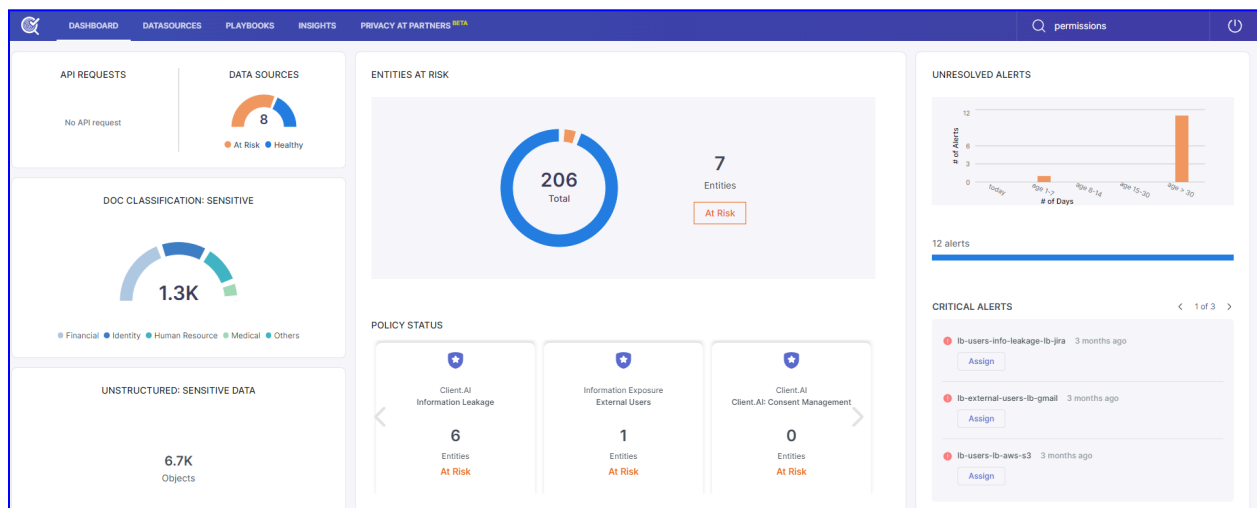
LightBeam Zero Trust Data Protection Platform

[LightBeam.ai](https://lightbeam.ai), the zero trust data protection pioneer, converges and simplifies data security, privacy, and AI governance, so businesses can accelerate their growth in new markets with speed and confidence.

Leveraging generative AI as a foundational technology, LightBeam ties together sensitive data **cataloging**, **control**, and **compliance** across structured, unstructured, and semi-structured data applications providing 360-visibility, risk remediation, and compliance for PCI, GLBA, GDPR, HIPAA among other regulations. Continuous monitoring with full data residency ensures ultimate zero-trust data protection. LightBeam is on a mission to create a secure privacy-first world.

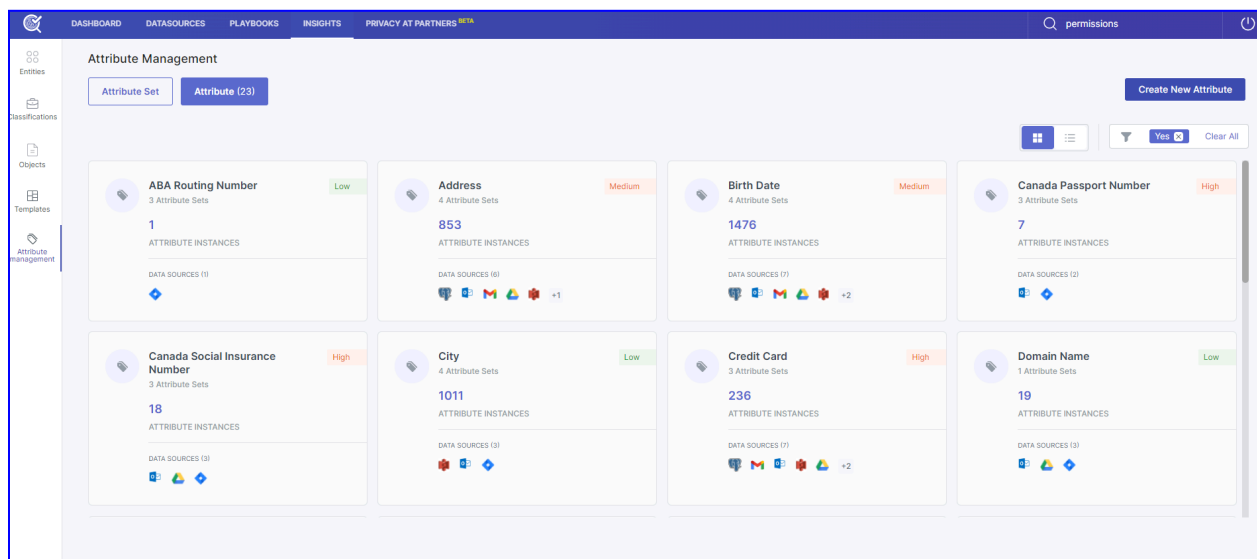
Main Dashboard

The main dashboard provides a high level view of all data sources where sensitive data is present, the entities (customers/employees/patients et al), and the attributes (sensitive data) that are being processed, and any policy alerts that need attention.



Attribute View

LightBeam has over 200 pre-configured sensitive attributes (sometimes called as elements, fields/columns) in it's system and is capable of recognizing their identifiers (sometimes called as values) from all the data sources; moreover users can also add their own proprietary attributes to the system and make it learn from the organizations various data sources. Attributes have 3 sensitivity levels based on their weight in the system (i.e. high, medium & low). Examples of attributes are U.S. Social Security Number, Loan Account Number, Medical Record Number and so on.



Entity View

Centered on the individual, the entity view provides a precise breakdown of what data is being held for any individual, and in what data sources it exists, and if there are any known associated risks. This view supports GDPR, CCPA, and other individual rights requests.

Name	Risk	# of Objects	# of Sensitive Attributes	Residency
jason flores	No	13	11	--
jacob christine hall	No	3	11	--
micheal brendan hughes	No	9	11	--
hannah cooper	No	16	8	--
craig johnson	Yes	4	11	--
Jackie Greene	No	3	11	--
michelle jorge brown	No	3	11	--
megan richard price	No	3	11	--
Craig Williams	No	3	11	--
Michael Zachary Welch	No	3	11	--

LightBeam Operational Phases

LightBeam’s Spectra DPA platform employs a three phase approach to managing privacy risk. These phases include Detect, Enforce, and Automate. Each of these phases builds on the previous phase to create a fully automated privacy management system that can;

1. Understand the existence and use of PI.
2. Create control policies with resulting actions.
3. Create automated tasks to execute control policies.

Detect

The initial LightBeam deployment step is to gain an understanding of the data environment. This includes connecting to applications and repositories to discover sensitive data elements called “attributes.” Attributes are contained in applications and repositories and are duplicated across the environment based on the relevant business processes. LightBeam uses API connections to analyze structured and unstructured

repositories and identify the data attributes, attribute types, and the related sensitivity levels. Then, an entity is resolved from the data related to an identified individual.

By understanding the data source and entity data that exists in the environment the LightBeam platform learns which data is important to an organization and its business processes. With this understanding as a foundation, LightBeam is then able to set policies as to how that data is stored, shared, and viewed.

During the detect phase LightBeam natively recognizes and classifies;

- 200+ common attributes including the common identifiers from a variety of countries.
- Industry attribute type sets like (Financial, Healthcare, Identity...)
- Unlimited client specific attributes - every LightBeam customer is unique and may carry sensitive data that is unique to their processing. LightBeam enables customers to add custom attributes. (e.g. Customer account number, employee number, member numbers, SKUs and other values)
- The classification of type and sensitivity of data contained in a document.
- Sensitive attributes detected across multiple data repositories which are linked using a machine learning algorithm to see if they belong to a single entity. The cross-linking of fragments of information to a central identity is a unique capability that helps customers understand not only if sensitive data is at risk but more importantly, whose data it is that might be at risk.
- The DETECT phase helps create data maps, RoPA reports and a 360-degree view of all information that's present about customers within an organization's systems.

Enforce

The enforce phase is used to establish the rules for data usage. There are four primary control components in the enforce phase. These include Policies, Permissions, Alerts, and Actions.

Policies

By mapping appropriate use of data for business functions, permitted data storage is identified and registered. Policies can be configured to respond to newly found data that deviates from permitted use. LightBeam Policies are configured to track both internal and external data sources. Each policy may contain multiple rule sets that define the search criteria and details about the data including; attribute sets and types, data sources, alert level settings and associated relevant regulations. Policies are configured via a query selection screen.

Policies screens include:

- The type of policies including; Internal, External and Leakage.
- The contact information for who an alert should be sent to.
- Setting of permissions list for white listing approved data sources.

Permissions

- Permission lists (also sometimes referred to as Permit Lists) establish and maintain an inventory of approved repositories and uses of PI.
 - Approved repositories are added to a permission list and used to compare new scan results against.
 - Alerts can be raised when a new instance is not found on a permission list.
 - Workflows are initiated by an alert to approve new instances and update the permission list so future findings will not raise an alert.

Alerts

- Alerts are used to notify system owners and others that a policy has been violated and that action may be needed.
 - Alerts are triggered based on rule sets inside policies.
 - Alerts can be set for specific applications or all connected applications.
 - Alerts can be set for specific attributes or individuals.
 - Alerts can trigger a workflow to drive a review and approve cycle.

Redaction

An example of automation is LightBeam's ability to redact or replace the presence of sensitive data within data repositories.

- The redaction control is used to maintain accurate processing records while reducing the use and risk of PI by removing select data from view.
 - Redaction control works on many document types including image files.
 - Redaction control can be set to replace or remove sensitive data while keeping the rest of the process information intact in LightBeam
 - Redaction filtering can be role based for eyes only operations.

In the example below a government ID is redacted of sensitive fields with some less sensitive information un-redacted. LightBeam can learn and scan images as well as documents and repositories.

The screenshot displays the LightBeam interface. On the left, the 'Entities' section shows email metadata: From: Aditya Ramesh, To: testinglightbeam@gmail.com, Sent: Aug 26, 2021 01:36, Subject: file. Below this, a 'Risk' indicator shows 'High' with 3 sensitive info items and 0 entities. A table below lists attributes with their confidence and mask status.

Context	Identifier	Attribute	Confidence	Mask
None	M*****g	Name	100%	On
None	!*****5	Birth Date	100%	On
None	g*****8	ID Number	100%	On

On the right, an attachment titled 'sample_jpg_(CC).jpg' is shown. It is a scanned image of a Government of India ID card. The card features the national emblem and text in Hindi and English. Sensitive fields like the name, birth date, and ID number are redacted with grey boxes. The card also includes a QR code and the slogan 'माझे आधार, माझी ओळख' (My Aadhaar, My Identity).

Guided by LightBeam established policies, the scanning engine, Spectra, continuously scans the data environment looking for changes to the data. New copies of files and uses can quickly be identified and either added to a permissions list, redacted or removed. By automating the execution of enforcement controls like alerting and redacting on a continual basis, an always-on accurate inventory of personal information is created. The process maintains an identity centric index that can be used to facilitate the retrieval of an Individual's PI and aid in the processing of Individual Rights Requests. Additionally DPA allows for duplicated data to be easily monitored and controlled reducing data leakage.

Data Security Automation for Data Warehouses

Automate

Utilizing LightBeams DSA technology to execute privacy process controls to continually scan, monitor and control data usage is a powerful tool for today's Privacy, Compliance, IT and IT Security teams. Automated monitoring of IT systems and information security controls has long been a part of most modern IT and security programs. Now the monitoring of sensitive data usage through automated processes greatly expands visibility, control and understanding of an organization's sensitive data use across the data lifecycle.

LightBeam customers can configure policies to trigger actions to manage the use of PI stored in data warehouses. LightBeam will identify any of the 200+ common attributes as well as any client custom attributes stored in structured and unstructured tables and columns. LightBeam will scan for PI and based on the rule set of the control policy that is in force to automatically take steps to;

- 1, Raise an alert and initiate a review and approve workflow
And / or
2. Replace or delete the PI stored in the dataset.

As an example A client may want to redact any sensitive information after a particular time frame. To accomplish this, Spectra would first analyze the data to see if it meets a date criteria. Select all records which are past a specified date, and that also contain sensitive information. The all sensitive data in the selected files would be deleted or replaced with a non sensitive value.

Data Warehouse Redaction

LightBeam can recognize sensitive data in Redshift and Snowflake data warehouses.

- By scanning and analyzing files and folders, sensitive data selected for redaction is identified.

- The use of selection criteria allows for attribute or data set specific redaction rule-sets to delete or replace the selected attributes.
- Files can safely be archived with all sensitive data redacted.

By establishing Policies, Rules sets, and Permission lists in LightBeam, automated monitoring not only provides visibility of the data inventory, it also continually enforces the data management rules to manage privacy risk in key applications.

By having a process to automatically remove PI after its retention period has expired, privacy officers have reduced their PI footprint and reduced their risk of inappropriate loss or disclosure.

Connecting Data Warehouses to LightBeam

Connecting any data warehouse data source to LightBeam is a simple 3 step process.

1. Create a new datasource instance.
2. Connect to the new data source.
3. Configure scan preferences.

Connecting Amazon Redshift to LightBeam

To begin from the LightBeam home dashboard, select DATA SOURCES. From the DataSource home screen select Add Data Source from the top right of the page. From the application list select New Redshift and continue with the steps below.

Step 1:

Complete the Basic Information page to create a new datasource. Keep the following guidance in mind when creating a new data source:

1. Make sure the instance name is not repeated, as it acts as metadata for the data source and must be unique. E.g., Redshift_HR_Sandbox.
2. Use the description to explain the kind of information the data source contains. E.g., All HR related documents stored here.
3. LightBeam uses the assigned owner email ID to send alert notifications.

4. Location tells where the data source is located.
5. Purpose tells for what purpose is this data source storing or processing data for.
6. Stage tells what stage the data belongs in could be source, processing, transactional, archival etc.
7. Use labels to select a label set for data tagging.
8. Enable source of truth if this is a gold source repository.

Redshift configuration

1 Basic information

Instance Name *

Description

Assign owner *

Source of Truth (SOT) ⓘ
 Mark this Data Source as 'Source of Truth'

2 Connection

Location

Purpose

Stage

Add labels

Step 2: Connection Details

1. Enter connection credentials.
2. Enter Host and port number information.
3. Set scanning frequency

Redshift configuration

1 Basic information 2 Connection

Username *

Password *

Host *

Port *

Frequency of scanning *

Step 3: Scan Settings

1. Select data bases for scanning

Redshift configuration

1 Basic information 2 Connection 3 Select database

Select database

Show all databases to select

Select specific database(s) that you have permission for

0 databases added for scanning

Database name

Connecting Snowflake to LightBeam

To begin from the LightBeam home dashboard, select DATA SOURCES. From the DataSource home screen select Add Data Source from the top right of the page. From the application list select New Snowflake and continue with the steps below.

Step 1:

Complete the Basic Information page to create a new datasource. Keep the following guidance in mind when creating a new data source:

1. Make sure the instance name is not repeated, as it acts as metadata for the data source and must be unique. E.g., Redshift_HR_Sandbox.
2. Use the description to explain the kind of information the data source contains. E.g., All HR related documents stored here.
3. LightBeam uses the assigned owner email ID to send alert notifications.
4. Location tells where the data source is located.
5. Purpose tells for what purpose is this data source storing or processing data for.
6. Stage tells what stage the data belongs in could be source, processing, transactional, archival etc.
7. Use labels to select a label set for data tagging.
8. Enable source of truth if this is a gold source repository.

Snowflake configuration

1 Basic information

2 Connection

Instance Name *
fw

Description

Assign owner *
fdw@gmail.com

Source of Truth (SOT) Mark this Data Source as 'Source of Truth'

Location
Select

Purpose
Select

Stage
Select

Add labels
Select labels

Step 2: Connection Details

1. Enter connection credentials.
2. Enter Host and port number information.
3. Set scanning frequency

Snowflake configuration


1 Basic information

2 Connection

Account name *

Username *

Password *

Warehouse *

Role

Step 3: Scan Settings

1. Select data bases for scanning

Snowflake configuration

1 Basic information 2 Connection 3 Select database

Select database

Show all databases to select

Select specific database(s) that you have permission for

0 databases added for scanning

Database name

Conclusion

Managing the appropriate use of personal information is challenging for any organization. Administrative controls like policies, procedures, and employee training are only as good as their execution which is often an afterthought in many organizations.

By using the power of AI to diligently scan and monitor data applications and repositories, data officers can apply significant technical control over how data is stored and processed. This in turn provides privacy teams new accurate pictures of their sensitive data and how it is used in the organization. By understanding all sensitive data in an organization LightBeams 360 degree view allows for more advanced reviews and proactive actions to be taken to manage privacy operations and reduce overall privacy risk.

About LightBeam

With its focus on Zero Trust Data Protection and Privacy Automation, LightBeam is pioneering a unique identity-centric and automation-first approach to the data privacy and data security markets. Unlike siloed solutions, LightBeam ties together sensitive data

discovery, cataloging, access, and data loss prevention (DLP), and makes the right identity-centric data available to the right people and teams. It becomes the privacy control tower providing a 360-degree view of PII/PHI sensitive data sprawl. LightBeam enables privacy officers to set policies to automate their enforcement, while information security executives can finally rest assured that sensitive data is being used and accessed securely.

Appendix

Revision History

Reference Architecture Update	Date	Author
First generally available version.	5/30/2024	Bill Schaumann