

**U.S. Securities and Exchange Commission**

---

**Enterprise Data Warehouse 2.0  
PRIVACY IMPACT ASSESSMENT (PIA)**



**March 20, 2022**

**Office of Information Technology**

# Privacy Impact Assessment

## Enterprise Data Warehouse (EDW) 2.0

### Section 1: System Overview

#### 1.1 Name of Project or System

Enterprise Data Warehouse 2.0

#### 1.2 Is the system internally or externally hosted?

- Internally Hosted (SEC)
- Externally Hosted
- (Contractor or other agency/organization) Cloud Environment

#### 1.3 Reason for completing PIA

- New project or system
- This is an existing system undergoing an update
- First developed: 6/10/2013
- Last updated: 2019
- Description of update: The update from EDW 1.0 to version 2.0 modernizes data analytics capabilities at the SEC by implementing a data lake (storage repository) and analytics environment in the cloud. Both systems are operating concurrently until full modernization is reached and EDW 1.0 is retired.

#### 1.4 Does the system or program employ any of the following technologies?

- Enterprise Data Warehouse (EDW)
- Social Media
- Mobile Application (or GPS)
- Cloud Computing Services
- Web Portal
- None of the Above

### Section 2: Authority and Purpose of Collection

#### 2.1 Describe the project and its purpose or function in the SEC's IT environment

The SEC maintains over 250 enterprise datasets with more than 100 datasets stored in the Enterprise Data Warehouse (EDW). EDW 2.0 implements a data lake (storage repository) and analytics environment in the cloud and consists of four areas of priority:

1. **Retire Hadoop, otherwise known as Enterprise Data Warehouse (EDW) 1.0** – Replace on premise EDW 1.0 infrastructure with a cloud environment with cluster computing capabilities. This gives SEC a more reliable working environment and removes the need for maintaining systems/software updates.
2. **Enable data access and security** – Establish an automated mechanism within EDW 2.0 that allows business owners to assign the appropriate level of access of data to users, down to the column level. This eliminates the need for the business owners to submit a ServiceNow request to manually add users to the datasets for which they need access.
3. **Deliver prioritized capabilities** – Fulfill requests by business owners for data analytic capabilities to improve execution of day-to-day work activities. Examples include standing up a scratch space and

# Privacy Impact Assessment

## Enterprise Data Warehouse (EDW) 2.0

notebooking capabilities, which provides data developers an area of temporary storage for their code and data to collaborate with peers and showcase their code and results.

### 2.2 What specific legal authorities, arrangements, and/or agreements allow the information to be collected?

EDW does not collect information directly from individuals and no new information is created on individuals. However, the underlying legal authorities and SORNs for the source data sets are identified in Section 2.4 below.

### 2.3 Does the project use, collect, or maintain Social Security numbers (SSNs)? *This includes truncated SSNs.*

- No  
 Yes

If yes, provide the purpose of collection:

If yes, provide the legal authority:

SSN is not collected by EDW. However, SSN may be contained in data stored in the repository from other systems and may be used by the end user as a part of their analysis based on a specific use case. EDW does not collect information directly from individuals and no new information is created on individuals. Therefore, the underlying SORNs for the various data sets provide the legal authority for the collection of the data.

### 2.4 Do you retrieve data in the system by using a personal identifier?

- No  
 Yes, a SORN is in progress  
 Yes, there is an existing SORN

The following SORNs are identified as applicable to EDW datasets:

- SEC-01 “SEC’s Division of Corporation Finance Records”
- SEC-02 “SEC’s Division of Investment Management Records”
- SEC-03 SEC’s Division of Trading and Markets Records
- SEC-05 “Office of Municipal Advisor Records”
- SEC-07 “Payroll, Attendance, Retirement, and Leave Records”
- SEC-10 “Personnel Management Employment and Staffing/Training Files”
- SEC-17 “Enforcement Files”
- SEC-25 “Information Pertaining or Relevant to SEC Registrants and Their Activities”
- SEC-26 “Mailing, Contact and Other Lists”
- SEC-29 “Tips, Complaints, and Referrals Records”
- GSA/GOVT- 7 “Personal Identity Verification Identity Management System (PIV IDMS)”
- OPM/GOVT-1 “General Personnel Records”
- OPM/GOVT-5 “Recruiting, Examining, and Placement Records”

### 2.5 Is the information covered by the Paperwork Reduction Act of 1995 (PRA)?

- No  
 Yes

# Privacy Impact Assessment

## Enterprise Data Warehouse (EDW) 2.0

### 2.6 Considering the purpose of the collection, what privacy risks were identified and how were those risks mitigated?

The privacy risk related to the collection include personal information is collected without a clear purpose or without clear legal authority. This risk is mitigated by collecting information as authorized and in accordance with the collection purpose identified in the underlying SORNs listed in Section 2.4.

### Section 3: Data Collection, Minimization, and Retention

#### 3.1 What information is collected, maintained, used, or disseminated about individuals? *Check all that apply.*

The system does not collect, maintain, use, or disseminate information about individuals.

#### Identifying Numbers

- |   |   |  |
|---|---|--|
| <input checked="" type="checkbox"/> Social Security Number                | <input checked="" type="checkbox"/> Alien Registration      | <input checked="" type="checkbox"/> Financial Accounts     |
| <input checked="" type="checkbox"/> Taxpayer ID                           | <input checked="" type="checkbox"/> Driver's License Number | <input checked="" type="checkbox"/> Financial Transactions |
| <input checked="" type="checkbox"/> Employee ID                           | <input checked="" type="checkbox"/> Passport Information    | <input checked="" type="checkbox"/> Vehicle Identifiers    |
| <input checked="" type="checkbox"/> File/Case ID                          | <input checked="" type="checkbox"/> Credit Card Number      | <input checked="" type="checkbox"/> Employer ID            |
| <input type="checkbox"/> Other: <a href="#">Click here to enter text.</a> |   |  |

#### General Personal Data

- |   |   |   |
|---|---|---|
| <input checked="" type="checkbox"/> Name                      | <input checked="" type="checkbox"/> Date of Birth     | <input checked="" type="checkbox"/> Marriage Records      |
| <input checked="" type="checkbox"/> Maiden Name               | <input checked="" type="checkbox"/> Place of Birth    | <input checked="" type="checkbox"/> Financial Information |
| <input checked="" type="checkbox"/> Alias                     | <input checked="" type="checkbox"/> Home Address      | <input type="checkbox"/> Medical Information              |
| <input checked="" type="checkbox"/> Gender                    | <input checked="" type="checkbox"/> Telephone Number  | <input checked="" type="checkbox"/> Military Service      |
| <input checked="" type="checkbox"/> Age                       | <input checked="" type="checkbox"/> Email Address     | <input checked="" type="checkbox"/> Mother's Maiden Name  |
| <input checked="" type="checkbox"/> Race/Ethnicity            | <input checked="" type="checkbox"/> Education Records | <input type="checkbox"/> Health Plan Numbers              |
| <input checked="" type="checkbox"/> Civil or Criminal History | <input checked="" type="checkbox"/> Zip Code          |   |
| <input type="checkbox"/> Other:                               |   |   |

#### Work-Related Data

- |  |  |   |
|--|--|---|
| <input checked="" type="checkbox"/> Occupation           | <input checked="" type="checkbox"/> Telephone Number           | <input checked="" type="checkbox"/> Salary              |
| <input checked="" type="checkbox"/> Job Title            | <input checked="" type="checkbox"/> Email Address              | <input checked="" type="checkbox"/> Work History        |
| <input checked="" type="checkbox"/> Work Address         | <input checked="" type="checkbox"/> Certificate/License Number | <input checked="" type="checkbox"/> Business Associates |
| <input checked="" type="checkbox"/> PIV Card Information | <input checked="" type="checkbox"/> Fax Number                 |   |
| <input type="checkbox"/> Other:                          |  |   |

#### Distinguishing Features/Biometrics

- |  |   |  |
|--|---|--|
| <input checked="" type="checkbox"/> Fingerprints | <input checked="" type="checkbox"/> Photographs | <input type="checkbox"/> Genetic Information |
| <input type="checkbox"/> Voice Recording         | <input type="checkbox"/> Video Recordings       | <input type="checkbox"/> Voice Signature     |
| <input type="checkbox"/> Other:                  |   |  |

#### System Administration/Audit Data

- |   |   |   |
|---|---|---|
| <input checked="" type="checkbox"/> User ID                               | <input checked="" type="checkbox"/> Date/Time of Access | <input checked="" type="checkbox"/> ID Files Accessed |
| <input checked="" type="checkbox"/> IP Address                            | <input checked="" type="checkbox"/> Queries Ran         | <input checked="" type="checkbox"/> Contents of Files |
| <input type="checkbox"/> Other: <a href="#">Click here to enter text.</a> |   |   |

#### 3.2 Why is the PII listed in Question 3.1 collected, used, shared, or maintained by the system or project?

PII is collected, used, shared, and maintained in EDW for business intelligence and analyzing data across disparate systems to identify patterns and trends which are difficult or impossible to discern via manual processes.

#### 3.3 Whose information may be collected, used, shared, or maintained by the system?

- SEC Employees

# Privacy Impact Assessment

## Enterprise Data Warehouse (EDW) 2.0

---

- Purpose: Employment and training data analysis
- SEC Federal Contractors
  - Purpose: Employment and training data analysis
- Interns
  - Purpose: Employment and training data analysis
- Members of the Public
  - Purpose: Data analysis of Registered entities submitting filings to the SEC
- Employee Family Members
  - Purpose:
- Former Employees
  - Purpose: Employment and training data analysis
- Job Applicants
  - Purpose: Employment and training data analysis
- Vendors
  - Purpose: Employment and training data analysis
- Other: Company executives, Reuters reporters, financial advisors
  - Purpose: Data analysis to enforce securities laws and regulations

### 3.4 Describe the PII minimizing mechanisms and if the PII from the system is being used for testing, training, and/or research efforts.

PII ingested from other internal SEC systems is minimized at the source collection point. Real public data (including PII) is used for testing purposes in the development and stage (lower) environments and PII is masked. In certain cases, EDW will use dummy data instead of real data to minimize the existence of PII in the lower environments. PII is not used for training or research purposes. EDW 2.0 has approved ATT-00042.

### 3.5 Has a retention schedule been established by the National Archives and Records Administration (NARA)?

- No.
- Yes. Official records are retained under DAA-266-2017-0002 "Data Marts and Warehouses"

### 3.6 What are the procedures for identification and disposition at the end of the retention period?

The Technical Architecture team keeps records of all data within the EDW 2.0 data lake until superseded by next scheduled update or when system is discontinued or obsolete, at which point the Technical Architectural team will delete the data from the corresponding Amazon S3 buckets.

### 3.7 Will the system monitor members of the public, employees, and/or contractors?

- N/A
- Members of the Public
  - Purpose:
- Employees
  - Purpose:
- Contractors
  - Purpose:

# Privacy Impact Assessment

## Enterprise Data Warehouse (EDW) 2.0

### 3.8 Considering the type of information collected, what privacy risks were identified and how were those risks mitigated?

There is risk of inadvertent or unauthorized disclosure of PII. This risk is mitigated by implementing access controls to limit access to those staff with a business need-to-know and by implementing encryption technology for stored and transmitted data.

## Section 4: Openness and Transparency

### 4.1 What forms of privacy notice were provided to the individuals prior to collection of data? *Check all that apply.*

- Privacy Act Statement
- System of Records Notice  
The SORNs listed in Section 2.4 are not provided to individuals prior to collection, but are published in the Federal Register and available on the SEC's website.
- Privacy Impact Assessment  
The EDW PIA is not provided to individuals prior to collection, but is available on the SEC's [website](#)  
Date of Last Update: 5/31/2020
- Web Privacy Policy
- Other notice:
- Notice was not provided.

### 4.2 Considering the method(s) of notice provided, what privacy risks were identified regarding adequate notice and how were those risks mitigated?

The primary privacy risk identified is that individuals may not be aware that their information is being collected in EDW 2.0 without an opportunity for them to consent to use of their information. This risk is mitigated by the publishing of privacy impact assessments for systems listed in Section 6.4. Also, there is minimal risk for adequate notice because the SORNs listed in section 2.4 and other notices identified in section 4.1 provide adequate notice to individuals.

## Section 5: Limits on Uses and Sharing of Information

### 5.1 What methods are used to analyze the data?

The following tools are used to analyze data in EDW:

- Starburst Presto - a distributed query engine that enables users familiar with SQL to execute queries across the data lake.
- AWS S3 - object storage service that serves as the main data repository for the data lake
- Amazon Elastic MapReduce (EMR) – cloud big-data platform for processing vast amounts of data using products such as Spark, Hive, and HBase
- Privacera – manages policies for access to the data lake
- Neo4j - graph database platform that used for artificial intelligence, fraud detection and recommendations

# Privacy Impact Assessment

## Enterprise Data Warehouse (EDW) 2.0

- Linkurious – used with Neo4j to provide graph visualization and analysis for fraud detection, anti-money laundering, intelligence, and cybersecurity.
- Business Intelligence Tools – MicroStrategy, Tableau, Power BI, etc.

### 5.2 Will internal organizations have access to the data?

- No  
 Yes

Organizations: All Commission Offices and Divisions, including Division of Economic and Risk Analysis (DERA), Division of Examinations (EXAMS), Division of Enforcement (ENF), Division of Trading and Markets (TM), Division of Investment Management (IM), Office of Information and Technology (OIT).

### 5.3 Describe the risk to privacy from internal sharing and describe how the risks are mitigated.

Information may be shared with SEC personnel who do not have a need-to-know the information in the course of their duties. To minimize this risk, EDW 2.0 uses Privacera (formerly Apache Ranger) application to automate data access and security based on a user's role in SEC's Active Directory.

### 5.4 Will external organizations have access to the data?

- No  
 Yes

Organizations:

### 5.5 Describe the risk to privacy from external sharing and describe how the risks are mitigated.

There is no risk to privacy from external sharing because data is not shared externally.

## Section 6: Data Quality and Integrity

### 6.1 Is the information collected directly from the individual or from another source?

- Directly from the individual.  
 Other source(s): EDW does not collect information directly from individuals. Information collected and stored in EDW is gathered from internal SEC source systems. Information is also provided from the commercially-available services providing public and social media data described in Section 2.1.

### 6.2 What methods will be used to collect the data?

EMR and AWS Glue, cloud-based platform and data integration tools, are mainly used to collect data. In addition, Python scripts on Amazon's Elastic Cloud Computing (EC2) platform may be used.

### 6.3 How will the data collected from individuals, or derived by the system, be checked for accuracy and completeness?

EDW does not perform data quality checks for accuracy and completeness but relies on the source system from which data is ingested to check for data quality.

### 6.4 Does the project or system process, or access, PII in any other SEC system?

- No

# Privacy Impact Assessment

## Enterprise Data Warehouse (EDW) 2.0

- Yes.  
System(s): EDW receives PII from various SEC source systems including BSS (FINRA Bluesheet), TCR, TRENDS, HUB, EDGAR, IARD, LEAP, PTCS, DERA FINRA DataMart, DERA data sets, Security Master Entity DB, Active Directory

### 6.5 Consider the sources of the data and methods of collection and discuss the privacy risk for this system related to data quality and integrity? How are these risks mitigated?

There is a risk that the SEC may receive outdated or inaccurate information. This risk is minimized because data integrity and quality control checks are performed by SEC source systems before data is ingested into EDW. However, data collected from social media is ingested as-is and not checked for data quality as discussed in section 6.3 above.

## Section 7: Individual Participation

### 7.1 What opportunities are available for individuals to consent to uses, decline to provide information, or opt out of the project? If no opportunities are available to consent, decline or opt out, please explain.

There is no opportunity for individuals using this system to opt out or decline to provide information because EDW does not directly collect from individuals. Opportunities to consent, decline, or opt out of EDW are addressed at the point of collection of the information and in accordance with the applicable SORNs for the source systems.

### 7.2 What procedures are in place to allow individuals to access their information?

Individuals seeking access to their information in a source system that provides information to EDW, may submit a request in writing to the FOIA/PA Officer, Securities and Exchange Commission, 100 F Street NE, Washington, DC 20549-2736 or email request to [foiapa@sec.gov](mailto:foiapa@sec.gov).

### 7.3 Can individuals amend information about themselves in the system? If so, how?

Individuals seeking to amend their information in a source system that provides information to EDW, cannot directly amend information about themselves in EDW. However, individuals may submit a request in writing to the FOIA/PA Officer, Securities and Exchange Commission, 100 F Street NE, Washington, DC 20549-2736 or email request to [foiapa@sec.gov](mailto:foiapa@sec.gov).

### 7.4 Discuss the privacy risks related to individual participation and redress? How were these risks mitigated?

The primary risks are lack of access to information and inability to seek redress and correction. This risk is mitigated by providing individual access or correction of the records as expressly permitted by the Privacy Act. Requests for privacy redress may be submitted to the FOIA/PA Officer, Securities and Exchange Commission, 100 F Street NE, Washington, DC 20549-2736 or emailed to [foiapa@sec.gov](mailto:foiapa@sec.gov).

## Section 8: Security

### 8.3 Can the system be accessed outside of a connected SEC network?

- No  
 Yes
- |   |                             |                              |   |
|---|-----------------------------|------------------------------|---|
| If yes, is secured authentication required? | <input type="checkbox"/> No | <input type="checkbox"/> Yes | <input type="checkbox"/> Not Applicable |
| Is the session encrypted?                   | <input type="checkbox"/> No | <input type="checkbox"/> Yes | <input type="checkbox"/> Not Applicable |

# Privacy Impact Assessment

## Enterprise Data Warehouse (EDW) 2.0

### 8.5 Does the project or system involve an online collection of personal data?

- No
  - Yes
- Public  
URL:

### 8.6 Does the site have a posted privacy notice?

- No
- Yes
- N/A

## Section 9: Accountability and Auditing

### 9.1 Describe what privacy training is provided to users, either general or specific to the system or project.

All SEC staff and contractors receive initial and annual privacy awareness training, which outlines roles and responsibilities for proper handling and protection of PII. SEC Rules of the Road ensure that employees and contractors are aware of their security responsibilities and how to fulfill them.

### 9.2 Does the system generate reports that contain information on individuals?

- No
- Yes

### 9.3 Do contracts for the system include Federal Acquisition Regulation (FAR) and other applicable clauses ensuring adherence to the privacy provisions and practices?

- No
- Yes
- This is not a contractor operated system

### 9.4 Does the system employ audit logging or event logging?

- No
- Yes

### 9.5 Given the sensitivity of the PII in the system, manner of use, and established safeguards, describe the expected residual risk related to access.

Although access to EDW is limited only to authorized SEC staff, the expected residual risk related to access to PII in the system can include the inadvertent handling or misuse of data. To mitigate this risk, user accounts for employees are synched with Active Directory and system privileges are granted based on defined roles.