Date of Approval: 10/09/2024
Questionnaire Number: 1571

Basic Information/Executive Summary

What is the name of your project (system, database, pilot, product, survey, social media site, etc.)?

Enterprise Data Platform - Neo4j

Acronym:

EDP-Neo4j

Business Unit

Information Technology

Preparer

For Official Use Only

Subject Matter Expert

For Official Use Only

Program Manager

For Official Use Only

Designated Executive Representative

For Official Use Only

Executive Sponsor

For Official Use Only

Executive Summary: Provide a clear and concise description of your project and how it will allow the IRS to achieve its mission.

Enterprise Data Platform (EDP) is designed as a universal data hub within Amazon Web Services (AWS) Gov Cloud (Treasury Cloud). It will ultimately contain source data from all major tax processing systems as well as supporting data necessary to conduct the IRS mission. Enterprise Data Platform (EDP) will serve as the main source of integrated taxpayer data for transactional and analytical needs. Enterprise Data Platform (EDP) will utilize Amazon Web Services (AWS) Databricks, Marketplace API, Redshift, MongoDB (Database) and third-party software such as Business Objects and Informatica. Enterprise Data Platform (EDP) will also integrate with IRS common services such as Negative Tax Id Number (NTIN). The data will only be accessible to authorized IRS personnel users that have gone through the Business Entitlement Access Request System (BEARS) approval process and will only be provided access for

their specific role. Neo4j is a graph database storing data as graph (Neo4j is not an acronym. It is the name). It helps to better understanding of the data, identify hidden patterns more easily. It supports various type of graph analysis like data profiling, schema analysis, node connectivity, triangle detection etc. on a vast amount of data and reveal hidden connections between the data helping IRS to find connected tax records for fraud identification. Refer to the attached Neo4j Graph Analysis Samples.docx word document showing various types of graph analysis that can be done with Neo4j. Neo4j can be used to analyze data collected by various applications and stored in databases like Databricks, Redshift, PostgreSQL and Oracle.

Personally Identifiable Information (PII)

Will this project use, collect, receive, display, store, maintain, or disseminate any type of Sensitive but Unclassified (SBU), Personally Identifiable Information (PII), or Federal Tax Information (FTI)?

Yes

Please explain in detail how this project uses sensitive data from inception to destruction (data lifecycle).

EDP receives sensitive data as-is from authorized IRS systems, and to that extent, the data is covered by the legal statutes already in place for collection and use of this data by IRS. EDP will inherit the mitigation/elimination processes in place by the IRS to eliminate the use sensitive data. Use of the data will be made available only to authorized business users with its corresponding BEARS entitlements. The duration for which this data will be stored on EDP, is dependent on the applicable data retention requirements established by Records and Information Management (RIM) and National Archives and records Administration (NARA), and these requirements will be met. Neo4j will be using data from EDP data source like Databricks, S3 Buckets for graph analysis.

Please select all types of Sensitive but Unclassified data (SBU)/Personally Identifiable Information (PII)/Federal Tax Information (FTI) that this project uses.

Address
Email Address
Employer Identification Number
Federal Tax Information (FTI)
Individual Taxpayer Identification Number (ITIN)
Name
Social Security Number (including masked or last four digits)
Standard Employee Identifier (SEID)
Vehicle Identification Number (VIN)

Cite the authority for collecting SBU/PII/FTI (including SSN if relevant).

PII for federal tax administration - generally IRC Sections 6001 6011 or 6012
SSN for tax returns and return information - IRC section 6109

Product Information (Questions)

1.1 Is this PCLIA a result of the Inflation Reduction Act (IRA)?

Yes

1.2 What is the IRA Initiative Number?

4.3: Improve Technology Operations | EDP Program number is PGM0001196

1.3 What type of project is this (system, project, application, database, pilot/proof of concept, power platform/visualization tool)?

Neo4j is a graph database management system which include a built-in database to store data in graph format and graphical visualization tool.

1.35 Is there a data dictionary for this system?

No

1.36 Explain in detail how PII and SBU data flow into, through and out of this system.

Enterprise Data Platform (EDP) Databricks is a cloud-based data platform holding consolidated IRS data from various IRS business units and, Simple Storage Service (S3) buckets are cloud data storage for IRS data. PII and SBU data in (EDP) Databricks, in S3 bucket or in other IRS business unit databases are read into Neo4j for graph analysis. After completing the analysis, the data in Neo4j can be removed and re-loaded if required for future analysis.

1.4 Is this a new system?

Yes

1.5 Is there a Privacy and Civil Liberties Impact Assessment (PCLIA) for this system?

Yes

1.6 What is the PCLIA number?

1498

1.7 What are the changes and why?

Installing Neo4j in EDP. Neo4j is a graph database for graph analysis on IRS data to identify hidden information between IRS tax records for fraud detection and forensic analysis.

1.8 If the system is on the As-Built-Architecture, what is the ABA ID of the system? If this PCLIA covers multiple applications shown on the ABA, please indicate the ABA ID for each application covered separated by a comma.

This system is built under Enterprise Data Platform, ABA id is 211416

1.9 What OneSDLC State is the system in (Allocation, Readiness, Execution)?

Execution

1.95 If this system has a parent system, what is the PCLIA Number of the parent system?

Enterprise Data Platform 8556

2.2 Please provide the full name of and acronym of the governance board or Executive Steering Committee (ESC) this system reports to.

Enterprise Services Governance Board (ESGB)

3.1 Does your project/system involve any use of artificial intelligence (AI), including virtual assistant, chat bot, and robotic process automation, as defined in Executive Order 13960?

No

3.3 Does this system use cloud computing?

Yes

3.31 Please identify the Cloud Service Provider (CSP), FedRAMP Package ID, and date of FedRAMP authorization.

Treasury Cloud/Workplace.gov Community Cloud FedRAMP High (TCloud/WC2-H) FR1801046750 03/02/2020

3.32 Who has access to the CSP audit data (IRS or 3rd party)?

Internal Revenue Service (IRS), Treasury Inspector General for Tax Administration (TIGTA) and General Accounting Office (GAO)

3.32 Does the CSP allow auditing?

Yes

3.33 Please indicate the background check level required for the CSP (None, Low, Moderate or High).

Moderate. Nobody in the team are given root access. System administrators are given only sudo root access. Sudo root access allows temporary permission to perform commands. Every time the Sudo user must re-authenticate before running commands.

3.4 Is there a breach/incident plan on file?

Yes

3.5 Does the data physically reside in systems located in the United States and its territories and is all access and support of this system performed from within the United States and its territories?

Yes

3.6 Does this system interact with the public through a web interface?

No

3.7 Describe the business process allowing an individual to access or correct their information.

The access and correction of taxpayer information are handled by the data source systems which are the IRS's system of record. EDP receives the information from the source systems. To that extent each source system ensures "due process" on information access, correction and redress. EDP, as the receiver of the data, inherits the compliance policies of that respective IRS data source system. Neo4j access entitlements are created in Business Entitlements Access Request System (BEARS) after working with Identity and Access Management team. Individuals submit request to Neo4j access entitlements via BEARS. That access request is reviewed and approved in 3 stages - 1) COR approval, 2) IRS Manager approval and 3) Neo4j team provisions the users access.

4.1 Who owns and operates the system (IRS Owned and Operated, IRS Owned and Contractor Operated, Contractor Owned and Operated)?

IRS Owned and Contractor Operated

4.2 If a contractor owns or operates the system, does the contractor use subcontractors?

Yes

4.3 What PII/SBU data does the subcontractor have access to?

Address, Email Address, Employer Identification Number, Federal Tax Information (FTI), Individual Taxpayer Identification Number (ITIN), Name, Social Security Number (including masked or last four digits) and Vehicle Identification Number (VIN). (FTI and ITIN require Live data waiver for all users).

4.5 Identify the roles and their access level to the PII data. For contractors, indicate whether their background investigation is complete or not.

Contractor/Subcontractor User: Read-only, access. Moderate background investigation level.

Contractor/Subcontractor Managers: Read-only, access. Moderate background investigation level.

Contractor/Subcontractor System Administrators: Administrator access; Moderate background investigation level.

Contractor/Subcontractor Developers: Read and write access. Moderate background investigation level.
Contractor/Subcontractor Testers: Read-only, access. Moderate background investigation level. Only contractors/subcontractors with completed background investigation will have access to SBU/PII data on EDP.
Neo4j Admin roles can view PII data. Background investigation for all contractors is complete. Users/Administrator that need to see PII data must submit Live Data Waiver by submitting 1) Treasury WC2 User Account_Access or Change Request Form 2-7 (copy attached) and 2) SBU 'Certification of Review' pdf form (copy attached).

4.51 How many records in the system are attributable to IRS Employees? Enter "Under 50,000", "50,000 to 100,000", "More than 100,000" or "Not Applicable".

More than 100,000

4.52 How many records in the system are attributable to contractors? Enter "Under 5,000", "5,000 to 10,000", "More than 10,000" or "Not Applicable".

More than 10,000

4.53 How many records in the system are attributable to members of the public? Enter "Under 5,000", "5,000 to 10,000", "More than 10,000" or "Not applicable".

More than 10,000

4.6 How is access to SBU/PII determined and by whom?

Only cleared individuals on a need-to-know basis with the corresponding Business Entitlements Access Request System (BEARS) entitlements and by following the SBU process (with permission granted per the requirements of Form 14664 - SBU Data Use Questionnaire and communicating with Privacy) will have access to SBU/PII.

5.1 Please describe any privacy risks, civil liberties and/or security risks identified for the system that need to be resolved and what is the mitigation plan?

Prisma Cloud and Orca Security are security platforms designed to secure applications and infrastructure in cloud. Security risks are identified on EDP by vulnerability scans using Prisma Cloud and Orca Security prior to and post-production. Any identified security vulnerabilities will be mitigated within the timeframes dictated by IRS Cyber.

5.11 Is there a Risk Assessment Form and Tool (RAFT) associated with this system on file with your organization or the IRS Risk Office.

No

5.2 Does this system use or plan to use SBU data in a non-production environment?

No

5.3 Please upload the Approved Email and one of the following SBU Data Use Forms, Questionnaire (F14664) or Request(F14665) or the approved Recertification (F14659). Select Yes to indicate that you will upload the Approval email and one of the SBU Data Use forms.

Yes

Interfaces

Interface Type

IRS Systems, file, or database

Agency Name

BMF - Business Master File database in Treasury T-Cloud Databricks

Incoming/Outgoing

Incoming (Receiving)

Transfer Method

Amazon Web Services Platform (AWS)

Interface Type

IRS Systems, file, or database

Agency Name

MTRDB - Modernized Tax Return Database (MTRDB) in Treasury T-Cloud Databricks

Incoming/Outgoing

Incoming (Receiving)

Transfer Method

Amazon Web Services Platform (AWS)

Interface Type

IRS Systems, file, or database

Agency Name

IRMF -Individual Returns Master File (IRMF) in Treasury TCloud Databricks

Incoming/Outgoing

Incoming (Receiving)

Transfer Method

Amazon Web Services Platform (AWS)

Systems of Records Notices (SORNs)

SORN Number & Name

IRS 26.019 - Taxpayer Delinquent Account Files

Describe the IRS use and relevance of this SORN.

Business and individual taxpayer return data will be used to serve the needs of the CFO office.

SORN Number & Name

IRS 24.030 - Customer Account Data Engine Individual Master File

Describe the IRS use and relevance of this SORN.

IRMF data will be used to support the individual taxpayer return needs of LB&I, EAD, PGLD Data Exchange, RRP, Direct File, BTA and IOLA.

SORN Number & Name

IRS 34.037 - Audit Trail and Security Records

Describe the IRS use and relevance of this SORN.

PII may be used for graph analysis purposes.

SORN Number & Name

IRS 26.020 - Taxpayer Delinquency Investigation Files

Describe the IRS use and relevance of this SORN.

Business and individual taxpayer return data will be used to serve the needs of the CFO office.

SORN Number & Name

IRS 24.046 - Customer Account Data Engine Business Master File

Describe the IRS use and relevance of this SORN.

BMF data will be used to serve the business return needs of LB&I, EAD, PGLD Data Exchange, RRP, Direct File, BTA and IOLA.

SORN Number & Name

IRS 22.062 - Electronic Filing Records

Describe the IRS use and relevance of this SORN.

Business and individual taxpayer return data will be used to serve the business return needs of LB&I, EAD, PGLD Data Exchange, RRP, Direct File, BTA and IOLA.

SORN Number & Name

IRS 00.001 - Correspondence Files and Correspondence Control Files

Describe the IRS use and relevance of this SORN.

Business and individual taxpayer return data will be used to serve the business return needs of LB&I, EAD, PGLD Data Exchange, RRP, Direct File, BTA and IOLA.

SORN Number & Name

IRS 37.006 - Correspondence, Miscellaneous Records, and Information Management Records

Describe the IRS use and relevance of this SORN.

Business and individual taxpayer return data will be used to serve the business return needs of LB&I, EAD, PGLD Data Exchange, RRP, Direct File, BTA and IOLA.

SORN Number & Name

IRS 22.061 - Information Return Master File

Describe the IRS use and relevance of this SORN.

IRMF data will be used to support the individual taxpayer return needs of LB&I, EAD, PGLD Data Exchange, RRP, Direct File, BTA and IOLA.

Records Retention

What is the Record Schedule System?

General Record Schedule (GRS)

What is the retention series title?

Transitory and intermediatory

What is the GRS/RCS Item Number?

5.2 item 20

What type of Records is this for?

Electronic

Please provide a brief description of the chosen GRS or RCS item.

EDP will receive BMF/IRMF extracts from IRS on-premises mainframe to support the individual and business return needs of LB&I, EAD, PGLD Data Exchange, RRP, Direct File, BTA and IOLA to be used for transactional and/or analytical purposes.

What is the disposition schedule?

All data meeting end of retention period requirements will be eliminated, overwritten, degaussed, and/or destroyed in accordance with National Archives and Records Administration (NARA)-approved disposition authorities for that system's data and done so in the most appropriate method based upon the type of storage media used in accordance with IRM 1.15.6.10.

What is the Record Schedule System?

General Record Schedule (GRS)

What is the retention series title?

Information Systems Security Records

What is the GRS/RCS Item Number?

3.2 Item 30 and 31

What type of Records is this for?

Electronic

Please provide a brief description of the chosen GRS or RCS item.

EDP will monitor and capture records as part of the user identification and authorization process to gain access to systems (System Access Records (AU 11)). Currently, access to EDP requires corresponding BEARS entitlements. *** EDP is in process of working with the IRS RIM office to determine and finalize retention schedules for data residing on the Platform.

What is the disposition schedule?

All data meeting end of retention period requirements will be eliminated, overwritten, degaussed, and/or destroyed in accordance with National Archives and Records Administration (NARA)-approved disposition authorities for that system's data and done so in the most appropriate method based upon the type of storage media used in accordance with IRM 1.15.6.10.

What is the Record Schedule System?

Record Control Schedule (RCS)

What is the retention series title?

Tax Administration Collection (Online Payment Agreement (OPA))

What is the GRS/RCS Item Number?

28 item 158

What type of Records is this for?

Electronic

Please provide a brief description of the chosen GRS or RCS item.

Direct Debit Installment Agreements (DDIA) (Form 433 Series) and related documents are records used by Compliance function taxpayer contact personnel to set up an agreement between the IRS and the taxpayer. The completed form permits the taxpayer to pay delinquent taxes through installment payments.

What is the disposition schedule?

Destroy immediately after 12 years. All data meeting end of retention period requirements will be eliminated, overwritten, degaussed, and/or destroyed in accordance with National Archives and Records Administration (NARA)-approved disposition

authorities for that system's data and done so in the most appropriate method based upon the type of storage media used in accordance with IRM 1.15.6.10

Data Locations

What type of site is this?

System

What is the name of the System?

Enterprise Data Platform

What is the sensitivity of the System?

Federal Tax Information (FTI)

Please provide a brief description of the System.

EDP Workplace Community Cloud (WC2) is designed as a universal data hub within AWS Gov Cloud (Treasury Cloud). It will ultimately contain source data from all major tax processing systems as well as supporting data necessary to conduct the IRS mission. EDP WC2 will serve as the main source of integrated taxpayer data for analytical needs. EDP WC2 will utilize Amazon Web Services Databricks, Marketplace API, Redshift and third-party software such as Informatica, Business Objects and Tableau. Authorized IRS personnel will use analytical tools such as Business Objects and Tableau to generate reports leveraging the taxpayer data. EDP WC2 will also integrate with IRS common services such as NTIN to prevent the unauthorized access to taxpayer data. EDP WC2 platform currently hosts multiple datasets - DDIA (Direct Debit Installment Agreement), CADE2 ODS (Individual Taxpayer Information), TAMS (Tax Account Management Services), MICA (Modernized Individual Custodial Accounting), BMF (Business Master File), IRMF (Individual Return Masterfile), Clean Energy (CE), and Direct File. Furthermore, new data APIs were deployed, enabling Read/Write Access to Business Taxpayer Account (BTA)and Clean Energy (Clean Vehicle Credits) users. All data on-boarded onto EDP will only be accessible to authorized IRS personnel users that have gone through the Business Entitlement Access Request System (BEARS) approval process and only be provided access for their specific role.

What are the incoming connections to this System?

DDIA (Direct Debit Installment Agreement), CADE2 ODS (Individual Taxpayer Information), TAMS (Tax Account Management Services), MICA (Modernized Individual Custodial Accounting), BMF (Business Master File), IRMF (Individual Return Masterfile), Clean Energy (CE), Direct File, and

Modernized Tax Return Database (MTRDB) via EFTU, Informatica IEP, or APIs.

What are the outgoing connections from this System?

DDIA (Direct Debit Installment Agreement), CADE2 ODS (Individual Taxpayer Information), TAMS (Tax Account Management Services), MICA (Modernized Individual Custodial Accounting), BMF (Business Master File), IRMF (Individual Return Masterfile), Clean Energy (CE) via EFTU, Informatica IEP, or APIs.

What type of site is this?

System

What is the name of the System?

Neo4j

What is the sensitivity of the System?

Personally Identifiable Information (PII) including Linkable Data

What is the URL of the item, if applicable?

prod.neo4j.edp.int.for.irs.gov

Please provide a brief description of the System.

Neo4j is a graph database storing data as graph. Neo4j allows one to analyze connected, high-volume and variably structured data assets. It helps to better understanding of the data, identify hidden patterns more easily. It supports various graph analysis like data profiling, schema analysis, node connectivity, triangle detection etc. on a vast amount of data and reveal hidden connections within individual data elements (nodes).

What are the incoming connections to this System?

The incoming connections are from S3 bucket and Databricks in Treasury T-Cloud for graph analysis of data in S3 bucket and Databricks.