Canada Data Retention Verification Tools_PRD

Product Manager	Shohei Kato
Document Status	DRAFT
Epic	CT05-204 - 2024 - Data Retention tooling enhancements BACKLOG
Target Release Date	April 30, 2024
Engineering Team	Shweta Bhale, Adam Farley, Kiwook Kwon, Brian Xu, Saurabh Kumar
Business Approval	

Purpose

This document outlines functionalities that are expected of Canada Data Retention Verification Tools based on requirements gathered by Canada Data Retention program (our stakeholders). In an effort to assist the team in successfully protecting personal information of our customers and compliance of the results, the tools automate verification that ensures the legitimacy of results processed by OneRemedy. Sign off of this document by Business team indicates the requirements listed below meet their business intent.

Introduction

Canada Data Retention Verification Tools automate verification process that ensures personal information that no longer services business needs have been remediated and compliance with Canada Data Retention Procedure

As outlined in the Canada Data Retention Procedures, Capital One Canada is required by Canada Branch Privacy Standard (5305.1.006) section 2.5 Limiting Use, Disclosure and Retention, to destroy or erase personal information (remediated) or make it anonymous once it is no longer required to fulfil the identified business purposes or to satisfy legal obligations in Canada. The Canada Data Retention Procedures details the specific requirements for the retention of personal information that was collected to support the Canada Card business. Failure of compliance results not only in regulatory penalties but loss of business brand and trust from our customers.

Every year, Canada Retention Program team conducts annual audit to ensure personal information that no longer serves business needs have been remediated accordingly and Canada Data Retention Verification Tools (Data Retention Tools) assists Capital One Canada in the pursuit. There are four possible remediation actions:

Retain: Data is retained without change.

Delete: The full dataset is deleted

Delete_in_scope: In scope records are deleted and the rest of the dataset is unchanged

Mask: Sensitive columns for in-scope records are masked

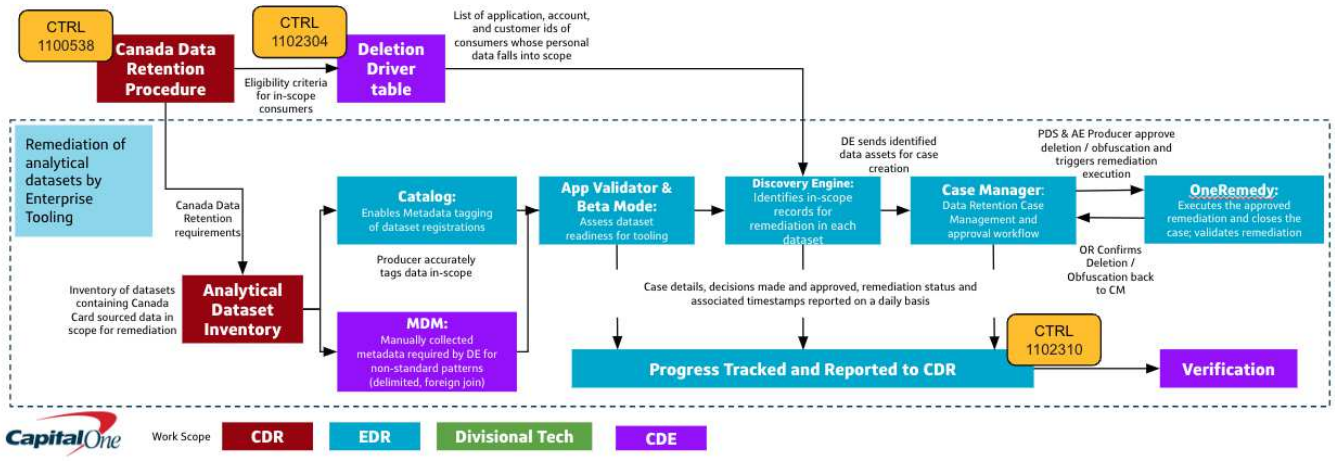
The data remediation is run by One Remedy (OR), an enterprise tool managed centrally in the U.S.. However, there were issues in the past where OR did not remediate the dataset at all which was only discovered during the manual verification by Canada Data Retention team. The manual verification process has been resource-intensive, prone to human errors, and limiting the scale of the verification operational. Our tools automate this verification process and this year alone, enabled the Canada team to conduct checks at ten times the previous capacity and with greater precision.

This product is built on the hypothesis that automation, scalability (aspirationally, 100% of all datasets), and speed are crucial for improving data verification processes. By automating these checks, our tools significantly reduce the risk of human error and ensures a more thorough and reliable verification process. It supports Capital One Canada in maintaining compliance with Canada Data Retention Procedure in a easy to use and navigate manner.

Background

This verification tools process the all datasets that go through Canada Data Retention program using OneRemedy, and there are two streams of work in Canada Data Retention Program: 1) Analytical Dataset by enterprise tooling and 2) Operational Data stores by divisional Tech Teams. This Canada Data Retention Verification Tools handle the datasets under the Analytical Dataset path, datasets (usually either in Snowflake or OneLake) that contain Canada Personally Identifiable Information (PII) are included. If the dataset resides in the platforms other than Snowflake or OneLake (e.g. S3 bucket, ElasticCache, ElasticSearch, metadata not stored in Catalog, etc...) the verification of the dataset undergoes the operational data store stream. The metadata of these datasets should be accurately stored in Catalog (the enterprise system) because Catalog is critical in remediation process in identifying in-scope datasets and the columns to be remediated by OneRemedy. Additionally, datasets that reside in the collab space are not included in the scope of these verification tools because their retention control is in place outside of the Analytical Dataset path (that involves OneRemedy).

Scope: Analytical Dataset Workflow



Once Canada Data Retention Procedure (audit process) is established, the account/application IDs eligible for retention program are stored in either Deletion Driver Table (if join keys are identified in Exchange, Driver table) or MetaData Management (also known as foreign key join, when join keys are not identified in Exchange, MDM). Driver Table and MDM are reviewed and published around August in each cycle. Once the target datasets are in the Case Manager, the remediation action for each target dataset as well as target columns is identified. Then it goes through the rest of the process including Discovery Engine (where in-scope records within dataset are identified), Case Manager (approves remediation and stores result from OneRemedy) and actually processing by OneRemedy. Canada Data Retention Tools' scope include 1) setting up Deletion Driver Table 2) setting up MDM 3) verifying remediation results by OneRemedy.

Target User

Our tools support Canada Retention Program and its participating members (including program sponsors and stakeholders such as PDS'). Our tools exist so that program members can conclude its annual audit and submit evidence with confidence. We assume fundamental understanding of the retention program and workflow because we do not assume training or education of the program.

Competition/Benchmark

We don't have competitors but the comparison is manual process or verification tools by US team can be used for benchmark.

Current State

As we complete productionization of the code at the time of writing this PRD (May 2024), we have productionized output table, switching from EID based access to batchID based processing, refactoring of code, and other requirements listed as V1 in *Feature Requirement section* below.

Success Criteria (OKR)

Canada Data Retention Tools aspire to make verification process easy, fast, reliable, and scalable for Canada Data Retention members

Metrics	Assumption	Frequency	How to Measure

Reach	**Canada Data Retention members can verify remediation results for all analytical datasets under CDR scope** Except for two cases listed below, the tools should be able to access all datasets 1) PHPD access is not granted 2) no view is available	Weekly Average	Number of datasets accessed and loaded by the tools (numerator) Number of datasets under CDR scope (denominator) **Target: 100% (aspirational)**
Turnaround	**Canada Data Retention members can verify remediation results within 24 hours after the dataset becomes eligible for verification:** **1) REMEDIATION_ACTION = REMEDIATION_SUCCESS** **2) CASE_STATUS = CLOSED** Measured in the number of hours since the dataset becomes ready for verification and verification result becomes available	Weekly Average	Difference between remediation date (OneRemedy/Case Manager) and Verification Date **Target: 24hrs**
Self-Sufficiency **(counter metrics)**	Number of times when Canada Retention Team requests for verification to our team as opposed to running ad-hoc verification themselves	Weekly Average	Anecdotal (e.g. slack message) Less than 2 incidents per month reported

Next Milestone

Milestone	Start Date	End Date	Notes/Comments
PoC/MVP development	Complete	Complete	
Production implementation	2024/03/06	2024/05/31	Please refer to this document for detailed implementation plan
PRD Document Sign Off	2024/03/15	2024/05/31	Version 1
Monitoring Phase	2024/05/02		

Requirements

Verification Tools Logic

RETENTION_ACTION	Description

DELETE	If a dataset has the following statuses • REMEDIATION_ACTION = REMEDIATION_SUCCESS • RETENTION_ACTION = DELETE Then CDR wants to make sure the target dataset has been physically deleted, not just by registration on Exchange. The verification is conducted in two steps because OneRemedy deletes the dataset 90 days after the deletion process is initiated: 1) The verification tools first check if it has been 90 days since remediated, if false, the tools give "DELETE_PENDING" result. By this stage, all access should be revoked for the target dataset. 2) 90 days after the access has been revoked (in the meantime, no new access request can be made to the dataset) via API, OneRemedy physically deletes the datasets. Our tools then check API of the respective data platform (i.e. Snowflake or OneLake) and give "PASS" to the datasets if it finds the datasets have been deleted. Our tools also check for registration on Exchange as second layer of verification. The result is kept in the output table but not shown in QuickSight Dashboard since the requirement by CDR is to have the target dataset physically deleted. ---- DELETE logic in detail
DELETE_IN_SCOPE	If a dataset has the following statuses: • REMEDIATION_ACTION = REMEDIATION_SUCCESS • RETENTION_ACTION = DELETE_IN_SCOPE then the verification tools count the number of unremediated records (i.e. CDE_TOTAL_UNREMEDIATED_RECORDS).
MASK	If a dataset has the following statuses: • REMEDIATION_ACTION = REMEDIATION_SUCCESS • RETENTION_ACTION = MASK then the verification tools count the numbers of unremediated records (i.e. CDE_TOTAL_UNREMEDIATED_RECORDS). If a dataset has more than 1 column under CDR scope, the tool uses the column with the largest volume of unremediated records to verify the remediation result.
RETAIN	Out of scope as OneRemedy takes no action on these datasets so there is nothing to verify

Feature Requirements

Category	Requirement	Version	Expected Scenarios
Reach	CDR wants the tools to conduct verification automatically as long as a dataset resides in either OneLake or Snowflake	V1	Our verification tools provide verification result based on the verification logic irregardless of platform, join, remediation action, data type. For example, A snowflake dataset populated through DirectWrite needs to be deleted. Our tool confirms no access to the dataset is granted as access has been revoked, and our tool gives "DELETE_PENDING" as result.
Reach	CDR wants the tools to conduct verification automatically for the following join types to identify remediation target records: Privacy Key Join (Deletion Driver Table), Foreign Key Join (MDM)	V1	
Reach	CDR wants the tools to conduct verification automatically for the following remediation actions: DELETE, DELETE_IN_SCOPE, MASK (including when a dataset contains multiple MASK columns)	V1	

Reach	For datasets reside in OneLake, CDR wants the tools to conduct verification automatically for the following file format: Parquet, CSV, AVRO, TXT	V1	At Post_90_day_confirmation, we recognize that the dataset has been physically deleted as well as registration on Exchange
Processing	CDR wants the tools to run on a scheduled basis every day at 7am without any manual interventions. Once the tools finish running, the results should be available in the dashboard almost instantly	V1	A CDR member logs in and checks verification process update daily
Processing	The tools should run on the following cadence: 1. When a case is first marked as ready to review. For this to be true, (1) REMEDIATION_ACTION = REMEDIATION_SUCCESS, (2) CASE_STATUS = CLOSED and (3) RETENTION_ACTION is one of DELETE, DELETE_IN_SCOPE, MASK in the Case Manager table. 2. If the case is marked as successfully verified and it is either a DELETE-IN-SCOPE or MASK case, the tool should do a follow up check after 30, 60, and 90 days. If it is a DELETE case, the tool will follow the DELETE timeline. If there are two Case ID assigned to a dataset ID, the verification tools ALWAYS pick the latest case ID for verification. If OneRemedy deems remediation is success but the verification tools assess otherwise (FAIL) on any run, the dataset is sent back to OneRemedy for re-processing. Depending on the circumstances, a new case ID may be assigned to the dataset in re-processing**. If a new case ID is assigned by Case Manager, the verification tool runs as prescribed above. If a new case ID is not assigned, then re-verification is done through manually triggering verification (as listed below as part of future enhancement plan).	V1	A Snowflake dataset whose data is populated through OneStream needs to be deleted in scope. Target records are identified through privacy join. The dataset fails first verification but passes a follow up verification. However, when verified again 35 days later, the dataset was repopulated and our tools failed the dataset. It passes a follow up verification and subsequently passes post_90_day_confirmation verification. ==== A Onelake dataset whose data is populated through DirectWrite needs to be masked. There are several records with empty/null values in the target columns. The dataset passes first verification but fails at Post_30_day_confirmation due to repopulation. After the dataset is reprocessed by OneRemedy and it passes a subsequent verification as well as post_90_day_confirmation verification.
Report	CDR wants to check verification results on QuickSight Dashboard to track overall progress and datasets that failed verification and follow up with OneRemedy. Please see here for detailed requirements such as columns, expected values, and sources.	V1	A CDR member logs in and checks verification process update daily at around 10 am, and downloads CSV file for downstream tasks.
Processing/ Report	CDR wants to see unmasked records for every column under CDR scope when the target dataset has RETENTION_ACTION = MASK and the target columns are tagged with Non-public Personal Information (NPI).	V1	A Snowflake dataset whose data is populated through DirectWrite has 4 columns that needs to be masked. Case Manager reports 200 non-compliant records and fully remediated but our tools identified the following records yet to be remediated. • Column A: 13 • Column B: 24 • Column C: 2 Verification tool should fail the dataset

Processing/ Report	CDR wants to track the datasets that are 1)RETENTION_ACTION = MASK and their columns are masked with single Y as opposed to YyYy so that CDR can work with PDS to investigate if the column indeed contains PII. A single Y masking means that the field length of the column is 1, which may not be PII. If indeed contains PII, then CDR can work with PDS to adjust the dataset accordingly. Depending on how target column is set up, there are two cases of single Y masking: 1. all values masked with single Y OR 2. mix of single Y and normal masking	V1	A CDR member finds a dataset masked with single Y masking pattern. S/he needs to instruct the PDS responsible for the dataset to make an adjustment on data format so that next time when OneRemedy processing the dataset again, the dataset is masked with proper YyYy format.
Maintenance	As CDE, we want the tools to run on the latest version of EMP in prod environment so that we can minimize potential security risk or service interruption from using an outdated version of EMP that is no longer supported.	V1	Our verification code is running on the latest version of EMP
Processing	As CDE, we want the tools to use batchID and app client_id to access EMP Serving, prod environment. If this is not implemented, our tools need to access the environment using EIDs (like EMP Training Space) and this method faces potential service interruptions. For example, if the member who EID the tools use to access the environment is away on vacation and our verification tools is interrupted.	V1	Batch ID and Client_ID are used to access EMP Serving environment
Processing	CDR wants the verification tools if the dataset is REMEDIATION_ACTION = MASK and contains multiple columns tagged as NPI.	V1	TBD
Processing	CDR wants to run a manually triggered verification in addition to the scheduled runs so that they can address urgent cases themselves. • At each ad-hoc run, new record is stored and show in QuickSight dashboard. By default, our tools keep the record of verification every time the tools run • Run book is provided as part of deliverable so our users can execute frictionlessly • Users can specify which datasets to run manually triggered verification	V2	TBD
Reach	CDR wants the tools to cover datasets that are joined by **Composite Key** (2 privacy keys are used to join and identify remediation target records) in addition to the current Privacy Key Join and Foreign Key Join to reach 100% coverage	V2	TBD
Reach	CDR wants the tools to cover datasets with **Delta Lake** data format (a data storage format making it easier to consume OL datasets)	V2	TBD
Reach	As CDE, we want the tool to handle datasets with **Schema Overwrite** issue When exchange and OneRemedy have different set of schemas for the target dataset, OneRemedy overwrites columns when remediated. It is hard for me to verify because the column names are different	V2	TBD
Reach	As CDE, we want the tool to handle datasets with **Conflicting Partition**, Partitioning column is inconsistent	V2	TBD

Reach	As CDE, we want the tool to handle datasets with **all data in one column**: All data is inserted in one column	V2	TBD
Inventory	Uploading datasets connected via foreign join keys is automated (MDM) automation • The tool will generate query script based on the inputs provided by PDS • CDR team can manually check if all target tables can be accessed and identify ones could not be accessed due to access issue	V2	TBD
Report	CDR wants to keep the results from previous cycles in a secure and easy to access manner **Requirements to be refined** However, past results do not necessarily have to be shown in QS Dashboard as QS Dashboard should track the progress for the current cycle.	V2	TBD
Processing	CDR team wants to automate the re-remediation process of datasets that once failed our verification from sending datasets back to OneRemedy CDR wants to run a manually triggered verification in addition to the scheduled runs so that they can address urgent cases themselves. • At each ad-hoc run, new record is stored and show in QuickSight dashboard. By default, our tools keep the record of verification every time the tools run • Run book is provided as part of deliverable so our users can execute frictionlessly • Users can specify which datasets to run manually triggered verification **Requirements to be refined**	Future Enhancement V3	TBD
Processing	CDR team wants to track the status of datasets that failed verification from 1) dataset waiting to be sent back, 2) they are sent back to One Remedy, 3) One Remedy reprocesses the datasets, 3) Case Manager updates REMEDIATION_ACTION = REMEDIATION_SUCCESS **Requirements to be refined** • Does CDR want to add new entry to the dashboard each time when the status is updated? • Does CDR want to include those details into newsletter/slack notification?	Future Enhancement V3	TBD

Processing	CDR wants to manually update datasets, based on business rules as an example yet to be reflected in them, to override previous verification results **Requirements to be refined** • Timing of the edit/manual updates: only • Do we allocate time limit for CDR to manually update data (e.g.30 days after verification gives "PASS")? • Should we allocate manual update to the entire datasets or only ones • How do we enforce governance? Can one person edit and update freely or require confirmation by others? • Should tools be re-run immediately after the edits are made?	Future Enhancement V3	TBD

Report	CDR wants to receive email/slack notification immediately after every verification tools run so that they do not have to go to QuickSight dashboard to get high level understanding of the verification status
	
	Requirements to be refined
	
	• Format: notification on Slack OR email
	• Contents:
	  ○ 1) number of datasets verified a) PASS b) FAIL c) ISSUE distribution (should masked_with_single_Y counted towards PASS?)
	  ○ 2) new datasets verified (difference from last week)
	  ○ 3) cumulative # of new datasets with PASS
	  ○ 4) completion % (against the entire CDR target dataset)
	  ○ etc...
		Future Enhancement
		
		V3
			TBD

Please find this document for detailed enhancement feature list

Scope Out

- Our tool does not identify and report the cause of how particular datasets have been repopulated
- Counting deleted-records is not technically feasible

Product Risk and Mitigation Plan

Risk	Follow Up Process / Mitigation Plan
Our tools may fail to pick up dataset that is ready for verification	Total number of verified dataset will be compared with the total number of remediated dataset and if the gap becomes large, CDE investigate the gap
Our tools cannot access the dataset	We'll identify dataset with access issues and CDR team can work with the enterprise tooling team to resolve them. For example, there are datasets that cannot be accessed due to PHDP access have not been granted, we'll provide a list of dataset IDs that cannot be accessed
EMP job failure - our verification logic is stored and ready for daily execution on EMP. However, sometimes EMP job could fail to execute	We'll set up a mitigation plan as well as Service Level Agreement where how long for issue detection and resolution. What course of actions will be taken
(Enhancement) Features related to coverage expansion cannot be implemented (e.g. schema overwrite, composite key, conflicting partition, all data in one column, delta lake, no access)	Except for ones we have no access, and as long as we can manually verify, we'll manually verify by picking a sample

EMP cannot process large files	We'll provide a list of dataset IDs that cannot be processed due to file size

=====

Footnote

* By principles, collab tables can only exist for 6 months without an approved extension. When the owners' of the tables request for an extension in Canada, they must confirm that they have no aged customer PII in the table before they are extended and extension cannot be more than an year.

** When a dataset needs to be re-processed, there are two paths 1) re-run the case through all channels which generates a new caseID after it was once closed and re-run is requested 2) dataset is re-opened to test repopulation logic at OneRemedy which uses the same caseID. The difference between these 2 is the process for what's needed to get the caseID. One requires the full utilization of all enterprise tools including a re-decisionning by the PDS/AEP and the other re-opens at One remedy only.