# Losses from Public Accounts

Databricks Example

## Executive Summary

This report is being prepared for KPMG after they have been hired in an audit capacity for the federal government. This external accounting firm is reviewing government data to offer advice on controlling expenditures and ensuring responsible use of public funds. In addition to reviewing the data in the traditional methods like line by line reconciliation of the balance sheet, the group is investigating whether to add a data lakehouse to their audit reporting process.

The goal of the data lakehouse would be to incorporate public data sets into the audit process effectively. This information could help focus a more detailed review in ministries or departments where the most problems may be occurring. It could also help to identify trends across time that would be hard to see in a ledger account level audit. The ideal design would support the type of data found in businesses as well as public sector agencies, so that it a similar design could be deployed in private sector audits as well.

The users of the data lakehouse would be the financial data analysts at KMPG performing the audit. The governmental clients could also be invited to use the dashboard area to have an interactive review of their own data outside traditional reports. The KPMG sales team plan to use a demo version of the dashboards in their pitch to win government audit reporting contracts, because interactive visualizations are a more engaging way to present financial data. If such a tool was used across multiple KPMG business units, interesting comparative information about customers in those markets could be found by more advanced analytics.

The data set chosen as the example for the data lakehouse is the “Losses of Public Money or Property as per the Public Accounts of Canada” located at: https://open.canada.ca/data/en/dataset/024ec008-653b-4632-ade8-3d6c0c27729a

This data set shows where the government must report losses in public accounts from property loss, benefits fraud, theft and all other type of loss. This data set was created in 2024 as an amalgamation of three separate data sets that reported similar information, “Losses of Public Property as per the Public Accounts of Canada”, “Losses of Revenues Due to Fraud or Willful Misrepresentation as per Public Accounts of Canada” and “Losses of Public Property as per the Public Accounts of Canada”. The field “Loss type” represents which of these three originating data sets the transaction is from. As this is the first version of this unified data set, there are a number of data quality and consistency issues that make the raw CSV difficult to work with. That made it a good choice for demonstrating the capabilities of the data lakehouse to process this information.

There are numerous other possible data sets to include. Essentially, financial data sets that are organized by department or ministry with a comparable time scale could be used. That might include:

- Expenditure information
    - Procurement data like contract award amounts or vendor service payments. Canada has had a number of public procurement scandals like the ArriveCan app and the Pheonix Payroll system. This should be an area of concern for the investigation of public losses. This is an example of that data: https://open.canada.ca/data/en/dataset/a1acb126-9ce8-40a9-b889-5da2b1dd20cb
    - Investment information like capital expenditures on equipment (This is a historical data set but the equivalent https://open.canada.ca/data/en/dataset/6d2f494b-30eb-45ef-b27e-7d1c23748fe1)
    - Anonymized payroll information. This could be used to investigate if there any link between public loss and executive compensation.
- Adding this type of expenditure information would allow reporting on loss as proportion of total operating costs
- Revenues
    - The government revenue data set does not have ministerial level detail currently: https://open.canada.ca/data/en/dataset/9d7364eb-e344-4908-b173-c151c1710b11
    - However a significant proportion of the losses in this data set relate to benefits fraud and CRA debt – what essentially starts as loss in expected revenue.
    - More timely and detailed revenue information could provide a unified overview to the financial data by including both sides of conventional accounting transactions

## Data Lakehouse Design

### Bronze layer

This layer will have one schema and one table. This table will contain the raw data from the CSV with columns for the date and time the data was ingested.

| Column Name | Data Type | Contents |
|---|---|---|
| **FSCL_YR** | String | Year loss reported |
| **YR_LOSS_RPTD** | String | For adjustments to prior years, year the loss was first reported. |
| **LOSS_TYPE_EN** | String | Loss type English. Denotes the originating data set of the transaction. |
| **TYPE_DE_PERTE_FR** | String | Loss type French |
| **INCDNT_TYPE_EN** | String | Incident type English. The majority of records are missing this field |
| **INCDNT_TYPE_FR** | String | Incident type French. The majority of records are missing this field |
| **MINC** | Int | Ministry numerical code. |
| **MINE** | String | Ministry name English |
| **MINF** | String | Ministry name French |
| **DepartmentNumber-Numéro-de-Ministère** | String | Deparrtment number |
| **DEPT_EN_DESC** | String | Department description English |
| **DEPT_FR_DESC** | String | Department description French |
| **INCDNT_EN_DESC** | String | Incident description English |
| **INCDNT_FR_DESC** | String | Incident description French |
| **TOT_INCDNT_NO** | Int | Total count of incidents |
| **CHRGD_CY_VT_NO** | String | Vote number, if any, against which the current year recovery charge was made |
| **TOT_LOSS_AMT** | Int | Total cost in dollars of loss |
| **TOT_CY_RCVRD_AMT** | Double | Total amount of loss recovered in current year. |
| **NOT_XPTD_RCVRD_AMT** | String | Amount of loss not expected to be recovered. |
| **C_XPTD_SY_RCVR_AMT** | Double | Amount of loss expected to be recovered in subsequent years. |
| **Ingested_timestamp** | Timestamp | Time that CSV was ingested |
| **Ingested_date** | Date | Date that CSV was ingested |

### Silver layer

This layer will have one table that has the same columns as the bronze layer. A number of cleaning steps will be applied to the data prior to writing records into the table.

Cleaning step #1: Drop any records with the YR_LOSS_RPTD  filled in. The reason for keeping only records with the YR_LOSS_RPTD column as blank is that otherwise they are an update to prior reported losses and could cause duplication. See this note published with the data set:

https://donnees-data.tpsgc-pwgsc.gc.ca/ba1/perteargentbienspublic-losspublicmoneyproperty/soutien-support-eng.html

Cleaning step #2: Standardizing incident type

This data has several data quality issues, the worst of which pertain to the incident type field. The following standardization will be performed:

| Original Incident Type | New Incident type |
|---|---|
| Cases before the Courts (Canada Revenue Agency's estimate of the amount of tax evaded or refunds fraudulently obtained) | CRA before the courts |
| Cases before the Courts (CRA's estimate of the amount of tax evaded or refunds fraudulently obtained) | CRA before the courts |
| Fraudulent use of customer's weather account | Offense or Other Illegal Act |
| Loss of revenues due to Customs Act infractions | Customs Act infraction |
| Loss of revenues due to Customs Act infractions - Non-report/Smuggling | Customs Act infraction |
| Loss of revenues dues to Customs Act infractions - Non report/Smuggling | Customs Act infraction |
| Loss of revenues due to erroneous credit card number | Offense or Other Illegal Act |

The biggest issue is that this field is null for the majority of records, even though the more verbose “Incident Description” fields are filled. Ideally, cleaning step #3 would read those fields and infer an incident type but that is outside of the scope of the current implementation.

## Gold layer

This layer will contain three tables, one with a broad set of data and two which provide summary statistics.

“Gold Full” table:

| Column Name | Data Type | Contents |
|---|---|---|
| **FSCL_YR** | String | Year loss reported |
| **LOSS_TYPE_EN** | String | Loss type English. Denotes the originating data set of the transaction. |
| **MINE** | String | Ministry name English |
| **TOT_INCDNT_NO** | Int | Total count of incidents |
| **TOT_LOSS_AMT** | Int | Total cost in dollars of loss |

“Loss by incident type” table:

| Column Name | Data Type | Contents |
|---|---|---|
| **INCDNT_TYPE_EN** | String | Loss type English. Denotes the originating data set of the transaction. |
| **TOT_LOSS_AMT** | Int | Total cost in dollars of loss |

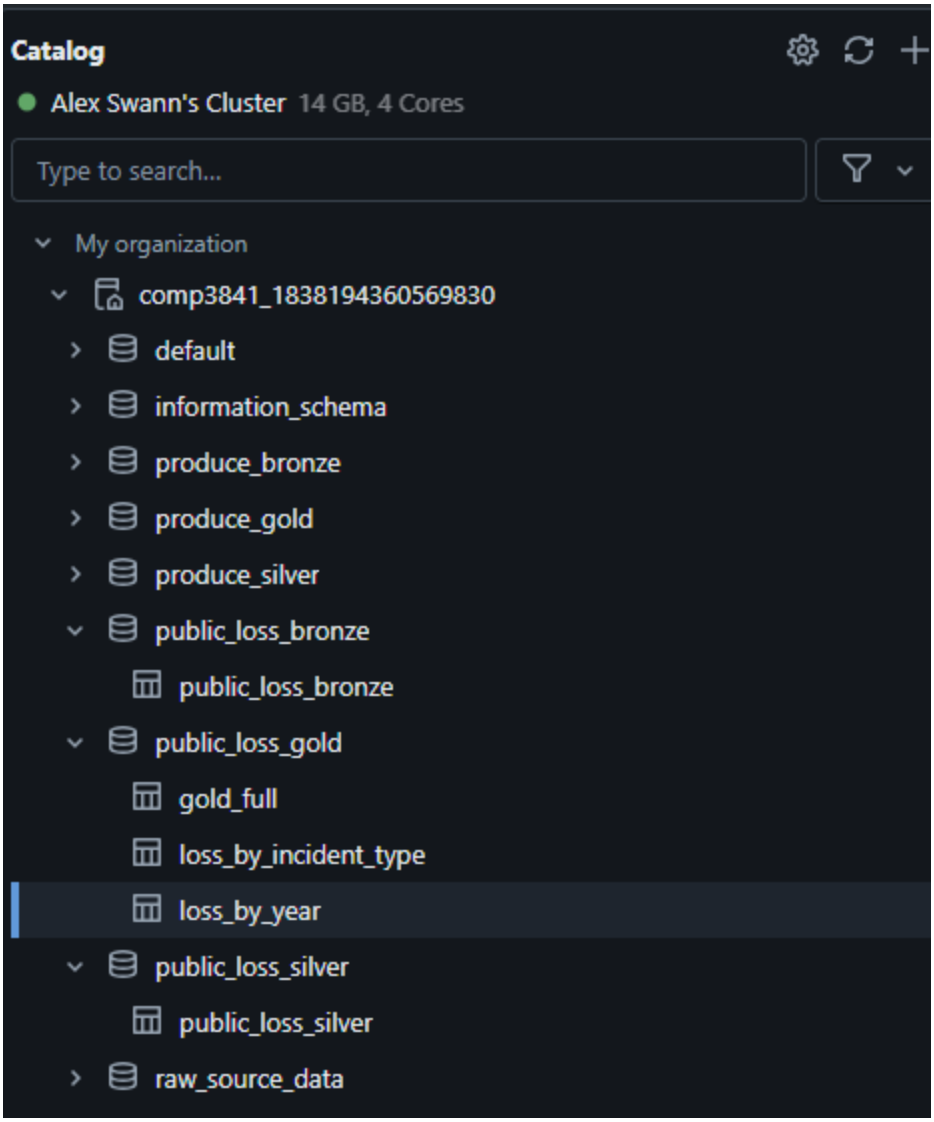
“Loss by year” table:

| Column Name | Data Type | Contents |
|---|---|---|
| **FSCL_YR** | String | Year loss reported |
| **TOT_LOSS_AMT** | Int | Total cost in dollars of loss |

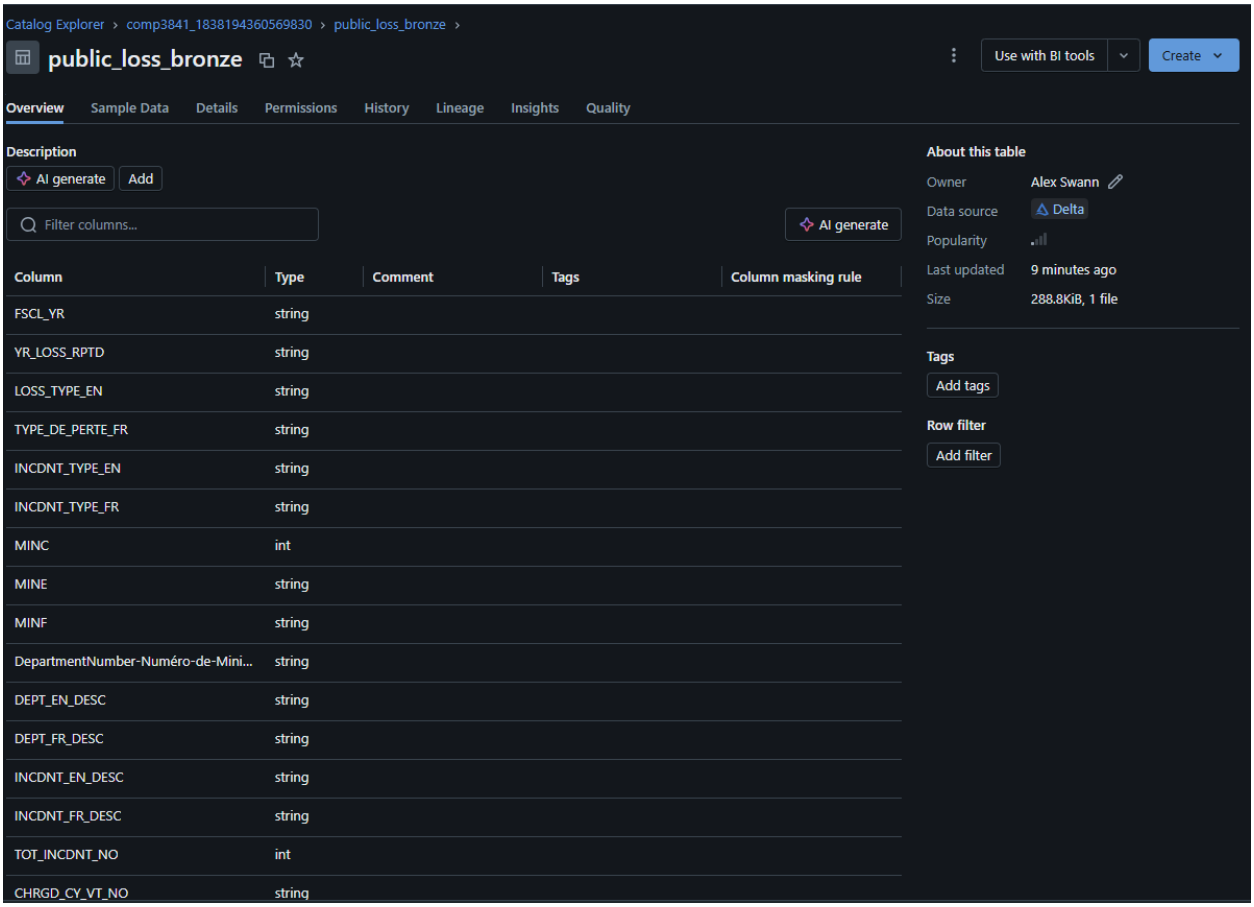
The gold tables are intended to demonstrate that either larger or more summarized data sets can be pulled from the larger amount of data depending on the type of user. Two dashboards will be created, one intended to show a high level summary of the data over time and another intended to be used with a filtered list of ministries. It’s intended to show that both static and interactive dashboards can be created.

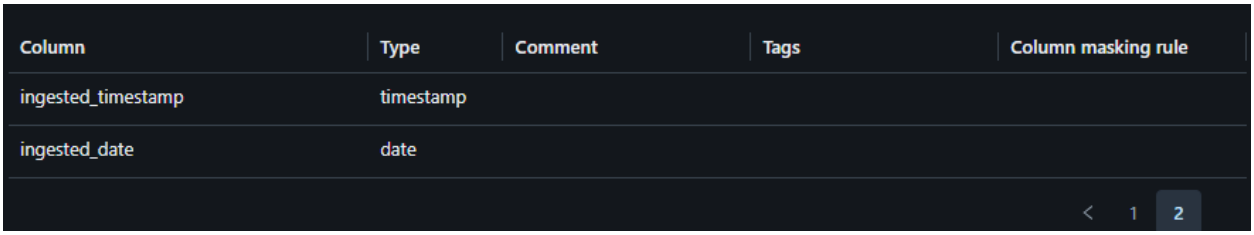
## Pipeline Implementation

Screenshots of schemas and tables:

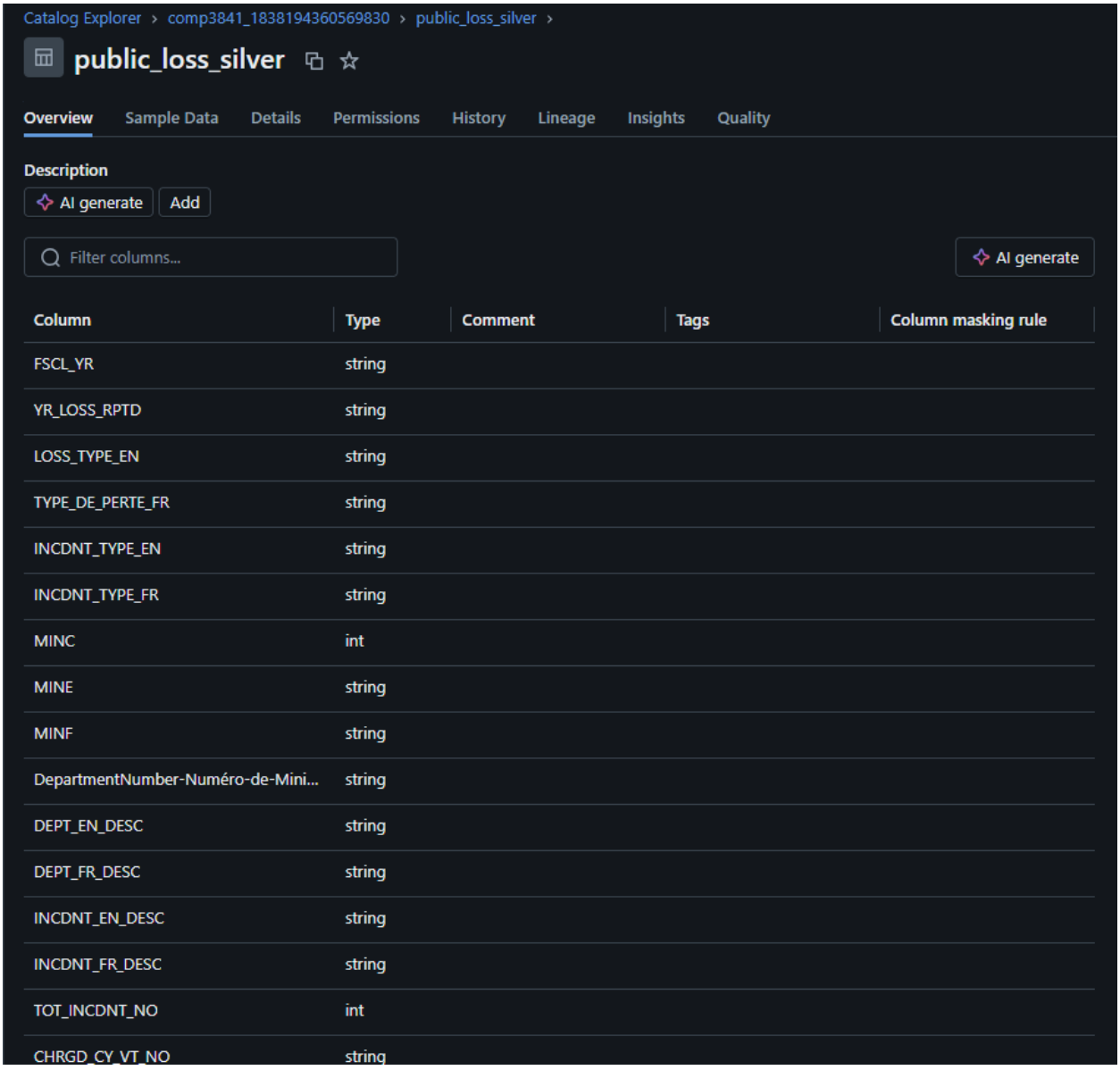



Bronze table:




Silver table:




| Column | Type | Comment | Tags | Column masking rule |
|---|---|---|---|---|
| ingested_timestamp | timestamp | | | |
| ingested_date | date | | | |

< 1 2

Gold tables:

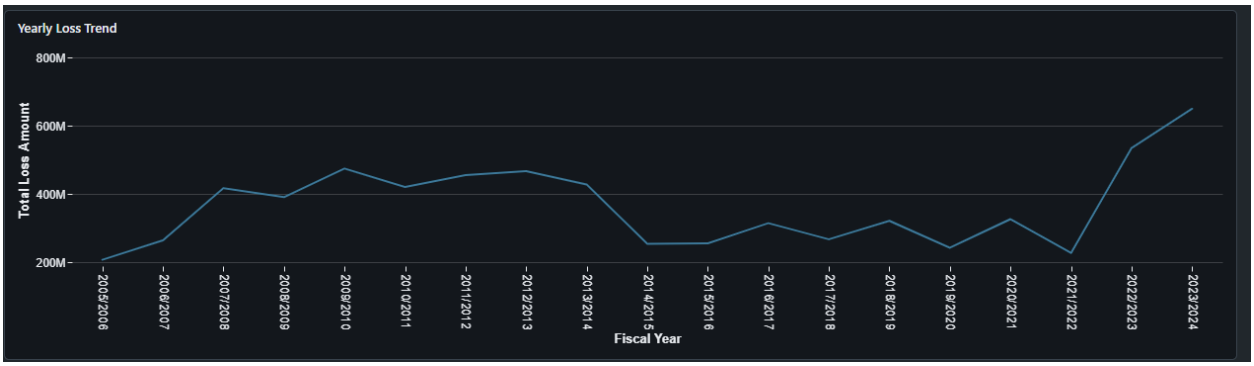
| Column | Type |
|---|---|
| FSCL_YR | string |
| MINE | string |
| LOSS_TYPE_EN | string |
| TOT_INCDNT_NO | int |
| TOT_LOSS_AMT | bigint |

| Column | Type |
|---|---|
| INCDNT_TYPE_EN | string |
| TOT_LOSS_AMT | bigint |

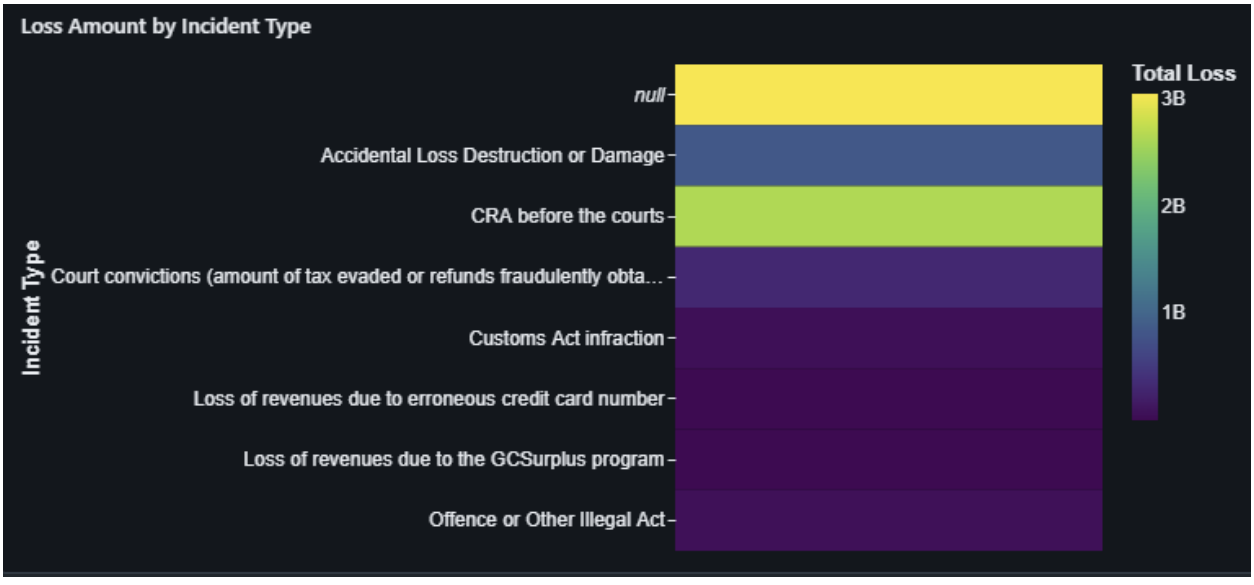
| Column | Type | Comment |
|---|---|---|
| FSCL_YR | string | |
| TOT_LOSS_AMT | bigint | |

## Dashboard Implementation

The first graphic shows the total loss amount over time using the “Loss by year” table. It’s intended to take the large data set and show the clear increase in total public loss amounts over time. The line graphic was chosen because it shows simplified information on one metric over a time period. This graphic would be used either by the compliance and audit team or by the government clients as an example of summary data:

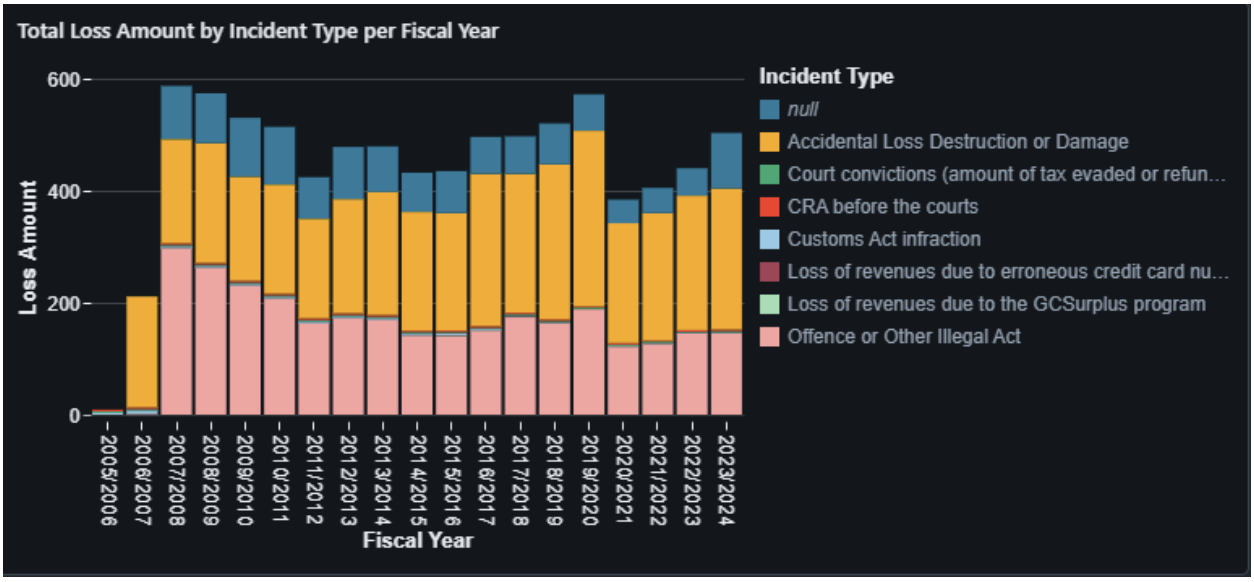



The second graphic shows the total loss amount by incident type using the “Loss by incident type” table. The most important part of this graphic is it demonstrates the large amount of “null” values in the incident type field. As this field is much easier to use than the incident description, this data quality issue limits the usability of this data set. The heatmap was selected because it shows proportion but is more readable than a pie chart. This graphic would be used either by the compliance and audit team or by the government clients as an example of summary data:

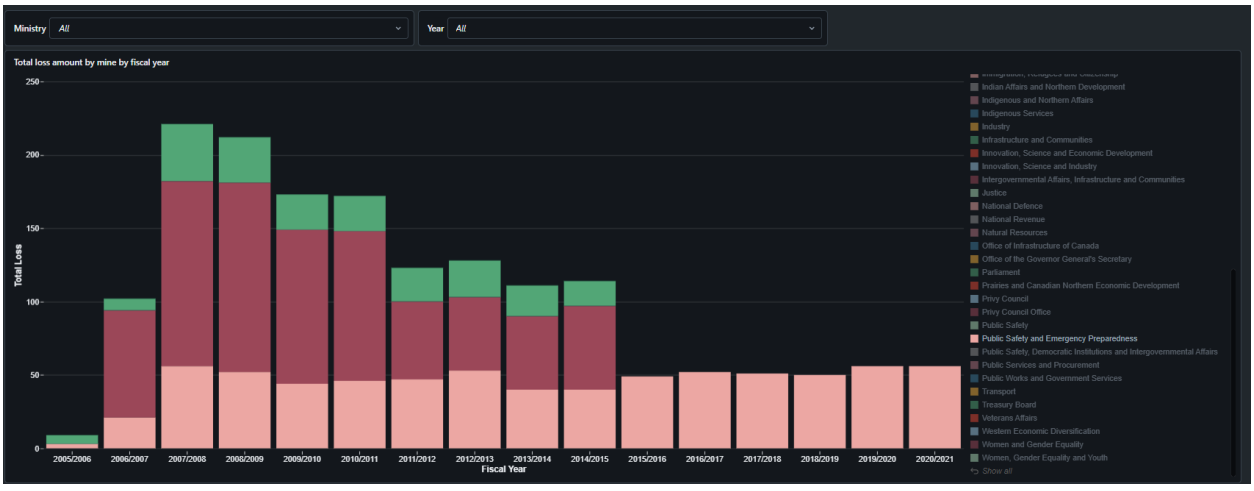



The third, fourth and fifth graphics use the “gold full” table. This table is more useful for visualizations because it allows for more flexibility in designing the graphics and their

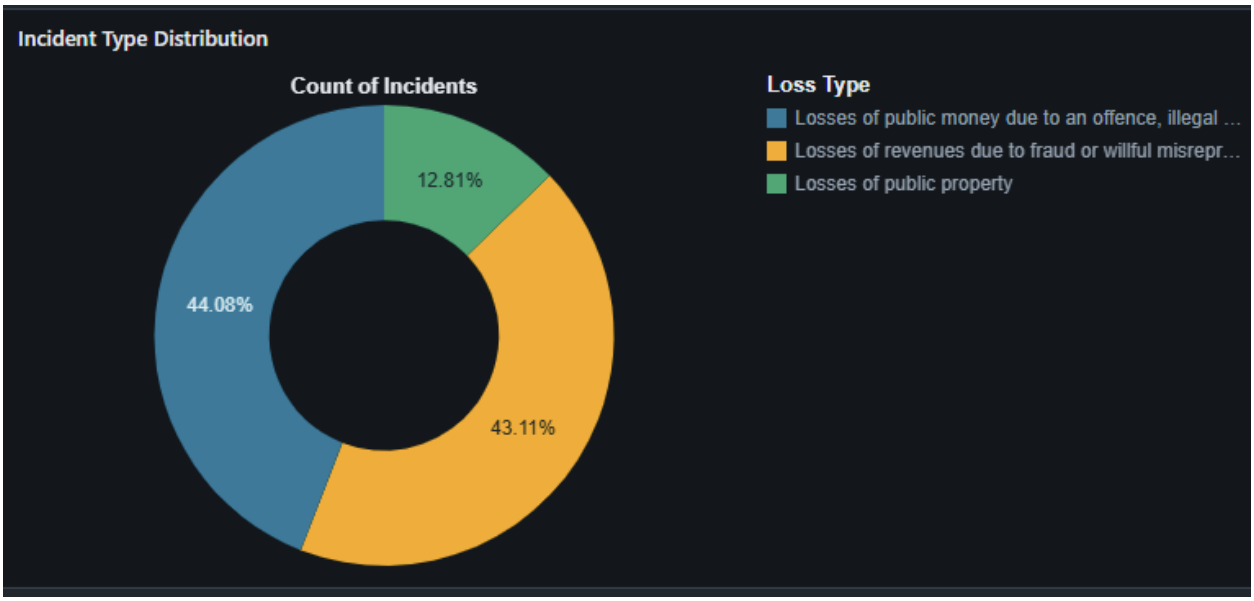
filters. The third graphic shows total loss by incident type over year as a stacked bar chart. The stacked bar chart is useful for showing time series data with multiple segments for comparison. This graphic would be used either by the compliance and audit team or by the government clients.




I made a second dashboard page and generated graphics that could be filtered by ministry or year. Without filters, the information from the “gold full” table is busy but with filters it can show useful comparison information between different ministries. It also revealed some further data inconsistencies in ministerial names once organized in this way. The stacked bar chart is useful for showing time series data with multiple segments for comparison. This information would be for the compliance and audit team or by the government clients. The interactive dashboard was also intended for use by the sales team.



The last graphic has count instead of cost information, as the number of incidents could reveal different details than the amounts. This is filterable by ministry and shows the amounts lost due to public property loss, fraud and other crime. The pie chart was selected because the number of categories is small. This graphic would be used either by the compliance and audit team or by the government clients. The interactive dashboard was also intended for use by the sales team.




## Pipeline Analysis

1. There are reasons to use a data warehouse versus a data lakehouse for this project but overall the flexibility of the data lakehouse may be preferable. The current data is not published very frequently, only annually. Traditionally, financial reporting occurs in a data warehouse and this is a type of financial reporting.

   However more detailed expenditure information is normally available on shorter timeframes, such as monthly or quarterly. While only this data set by itself could likely be housed successfully in a data warehouse, the inclusion of more diverse data updated more frequently could make a data lakehouse the more appropriate tool. Fraud detection is more commonly done in a data lakehouse type deployment, and ideally this would be a set of fraud investigation tools.

2. The Public Loss data set is published annually and so is the revenue, contract awards and capital investment data sets. Ideally, either monthly or quarterly expenditure information could be added so that more proactive review of ongoing spending should as vendor services for procurement could be monitored.

3. Batch data updates are scheduled and latency is acceptable, whereas streaming updates cause the data warehouse to be updated with transaction data almost in real-time. It is more likely for batch data to be consumed in large amounts, and more likely for streaming data to have new transactions continuously added.

   Within Databricks, you might use a “External volume” type Volume for the streaming data, as then Databricks could ingest data from a specified location on a more frequent basis. For both scenarios you would want to use Notebooks to automate the extraction, transforming and loading process, but the workflows associated with the streaming data would have shorter timeframes.