ADASTRA

Data Best Practices

Final Report

December 2022

REGION OF DURHAM

Introduction

The Region of Durham engaged Adastra Group to implement Microsoft Purview to enable insights into the Region's existing data systems and related services. The project success enabled savings and efficiencies through data insights and make data driven decisions.

The goals of the project included:

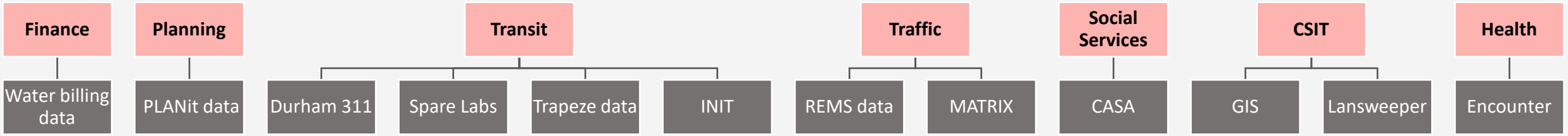
- Developing a model to classify data resources listing the data systems, high level data ownership, business process/service relationships and common corporate data sets;
- Developing a diagram to show the relationships between data sources and services;
- Recommendations for quick-win targets for data sets to be brought into Purview;
- Identify opportunities for system integrations and process automations, enabling cost savings and improved processes
- Identify potential duplication of data, barriers and constraints.

This report will provide an overview of the project, recommendations, next steps, and a breakdown of the identified potential cost savings and efficiencies.

Data Catalog Organization

The Region of Durham had access to Microsoft Purview through its Microsoft Azure tenant and Microsoft 365 enterprise licensing. To facilitate this project Microsoft Purview was used to scan, classify, set up glossary terms, apply lineage, label sensitive data, and explore data loss prevention of a select number of data sources. This work would provide the baseline for the identification of cost savings, efficiencies, and recommendations.

With the installation of Purview, we have included twelve areas into the scope for the project, including disparate areas with common information for insight gathering. The areas and data sources included are the following:



Glossary

For each of the subject areas, we engaged Subject Matter Experts and Data Stewards to ensure quality classification and glossary of terms for the dataset. In associating the glossary of terms, we employed a standard template for term consistency, quality and accuracy. This will be part of operational processes to support additional database scans.

Sample Glossary Terms

Nick Name	Name	Definition	Experts	Stewards
ARN	Region Of Durham_CSIT_GIS_ARN	The Assessment Roll Number of the associated parcel where the address is located. A 20-digit ARN will only occur when there is a condominium complex, all other ARNs will be 15 digits.	Subject Matter Expert	Subject Matter Expert
BIRTH_DAY	Region Of Durham_CSIT_GIS_BIRTH_DAY	The day of the month that the individual was born	Subject Matter Expert	Subject Matter Expert
Property Identification Number	Region Of Durham_Common Terms_Property Identification Number	Property Identification Number for land parcels in land registry system	Subject Matter Expert	Subject Matter Expert
Common Terms	Region Of Durham_Common Terms	Terms which applies across Business units will be maintained under this Term to avoid duplication while business unit specific terms will be maintained under relevant Business units Term	Subject Matter Expert	Subject Matter Expert
Assessment Roll Number	Region Of Durham_Common Terms_Assessment Roll Number	Assessment Roll Number for property parcels in MPAC assessment data	Subject Matter Expert	Subject Matter Expert

Classification

We focused on the areas across the organization, where different stakeholders would be able to gain insights into their datasets. The assignment of attributes to system or custom classifications, enabled alignment to known business terms. In doing this classification, we were able to determine and define common terms across the organization.

Sample Classification

	Classification Rule		
Classification Name	Column Name	Data Pattern	Dictionary
PostalCode	PostalCode PCode	A9A9A9 A9A 9A9	
Canadian Province	Canadian Province		for list of values please refer to 'CanadianProvinces' worksheet
ApplicationStatus	ApplicationStatus		for list of values please refer to 'ApplicationStatus' worksheet
MaritalStatus	MaritalStatus		for list of values please refer to 'MaritalStatus' worksheet
Municipality	MUNICIPALITY MUNI TOWN		Look up table on Municipality sheet
City	CITY TOWN		Look up table on CitiesTowns sheet
SamId	SamId SId SamsId	999999999	
ApplicationNumber		AA99-9999	

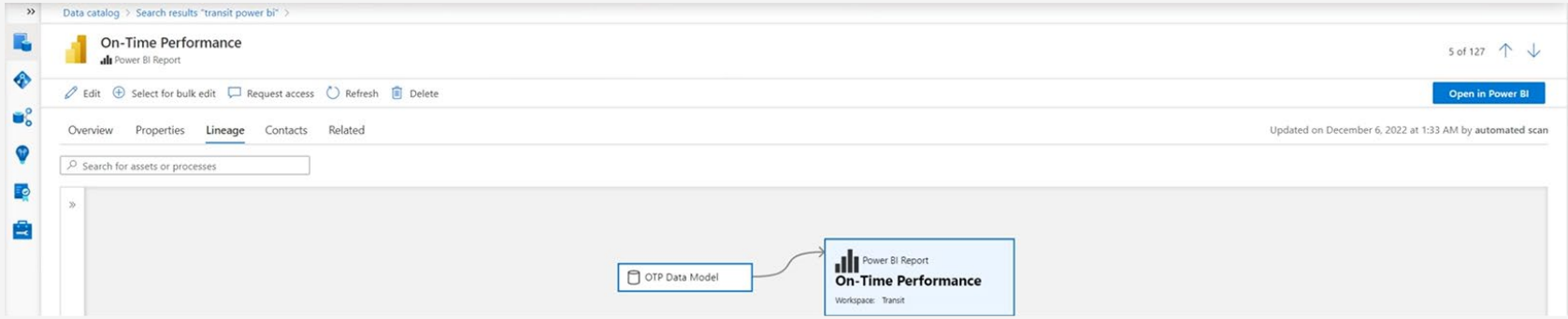
Data Lineage

Data lineage is the process of describing what data exists, where it is stored and how it flows between systems. One of the features of Microsoft Purview is the ability to show the lineage between datasets created by data processes. Metadata collected in Microsoft Purview from enterprise data systems are brought together to show an end-to-end data lineage. Each system supports a different level of lineage scope.

As part of the Data Catalog project, we identified in-scope processes and mapped out their Data Lineage. This provides insights into the organization's data flow from Source to Consumption Layer. This will provide time savings for support and maintenance, by providing clear insights for tracking process flows and identifying impacted processes and decision making.

Benefits of mapping Data Lineage:

- Capture the changes and where the data has resided through the data life cycle.
- Track data in reports.
- Impact analysis.

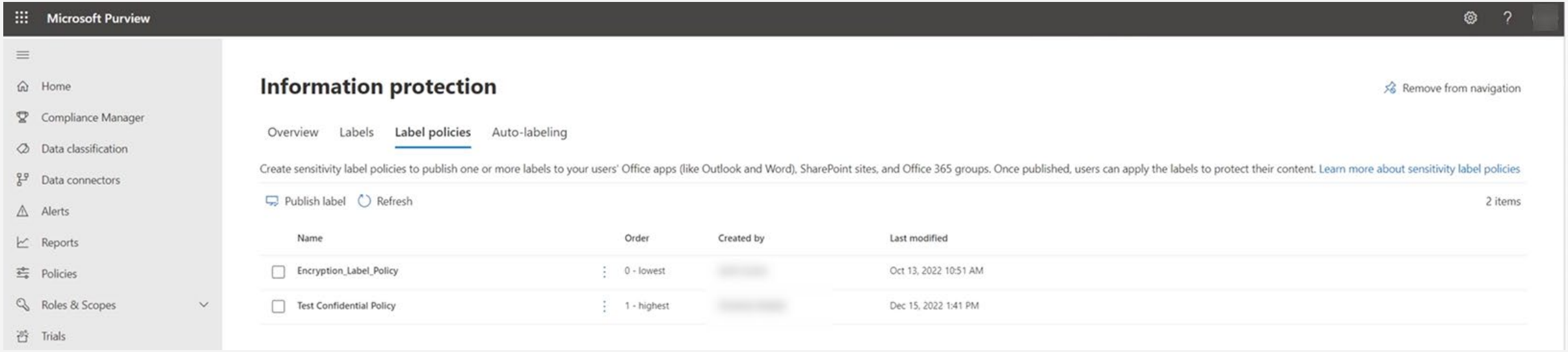


Sensitivity Labeling

Sensitivity labels are used to classify and protect content. They are the digital equivalent of ‘stamping’ a document to share how information can be used, shared or viewed. Applying sensitivity labels to your content enables you to keep your data secure by stating how sensitive certain data is in your organization. Awareness of the sensitivity will lead to reduced data loss and increase data security awareness.

We applied the three sensitivity labels to a defined SharePoint site to demonstrate the feature and benefits through;

- Creating a listing of Sensitivity Labels; Public, Internal and Confidential.
- Defining the scope for the Labeling.
- Providing recommendations for users to see the Labels and potential Data Loss Prevention measures.



Data Loss Prevention

- To help protect the organization’s sensitive data and reduce risk, we need a way to prevent users from inappropriately sharing it with people who shouldn't have it. This practice is called data loss prevention (DLP). In Microsoft Purview, we implemented data loss prevention by defining and applying DLP policies. The DLP policies will directly prevent data loss, breaches and improve awareness of data literacy. With a DLP policy, you can identify, monitor, and automatically protect sensitive items by:
- Applying policies to detect risky behavior and unauthorized sharing / breaches
- Designing policies for DLP, define control objectives
- Identifying DLP templates for use
- Identifying scope for workloads of DLP

Project Outcomes

This project provided the Region with a foundation to better understand their data assets and the services that use the data.

The project has provided the Region with the following outcomes and identified challenges:

Outcomes	Challenges
A structure to organize the Region's data sources and tables.	Identified some challenges with defining business glossary terms, as these values did not currently exist.
Classified the data in an automated way.	Identified datasets with common and/or duplicate data.
Traced the lineage data from data source through to reporting layer.	Identified the lack of accurate and complete metadata across the organization.
Provided a framework and single source of truth for business terms.	Identified that data literacy needs to be championed and improved across the organization.
Introduced a method to label data assets based on its sensitivity.	Identified that data silos exist and building a data catalog can make them discoverable and lead to sources of knowledge.
Introduced a method to notify risky behaviour and prevent data loss.	
Realized cumulative benefit through governance to classify, label and apply document Lineage, to assist in the prevention of data loss and data breaches.	

Recommendations

The implementation of Purview onto the Microsoft platform allowed the Region to increase the understanding of the in-scope data assets through data governance. This was done through the capture of common Business Terms, Classification of data attributes, building Data Lineage of the processes, assigning Sensitivity Labels and implementation of DLP practices to safeguard the Region's data.

It is recommended to expand the governance processes to more divisions and engage more staff to leverage the learnings for data insights and governance. The experiences of the project team, through limited scope, will allow for education and expansion of governance to be completed in dedicated blocks that will link up with completed areas, with an end goal of enterprise adoption. This will allow the identified savings to be realized across the organization.

As part of the initiative, the project team learned more about Data Governance Best Practices and industry standards that the Region should be developing. This process identified a need for accurate documentation and procedures for data governance, to secure the digital data. We have identified a data literacy gap, and a need to educate staff across the Region. The roll-out of improved documentation will improve efficiencies in gathering, sharing, using and managing information, throughout the Region.

In assigning Classifications and Glossary Term definitions, there is a need to identify and assign Data Owners and Data Stewards for data across the Region. The assignment of Data Stewards is key to maintaining and enforcing accurate definitions and usage of data. Through this tool and education of staff, there will be stronger compliance and security in the use of the data. In having consistent and knowledgeable Data Stewards, there will be a more consistent application of terms, higher frequency of accurate reporting and fewer areas of reporting inconsistencies.

With additional effort, the Purview insights can be expanded to span across databases to provide sensitivity labels for encryption, access restrictions, and visual markings. These can be considered for future project deliverables.

Recommendations (cont'd)

It is recommended that the Region build on the Data Governance policies to better manage digital data, by considering the following governance concepts:

1. Classification is the process of organizing data into logical categories that make the data easy to retrieve, sort, and identify for future use.
2. Business glossary term defines the business vocabulary for an organization and helps in bridging the gap between various departments in your organization.
3. Sensitivity labels are a type of annotation that allows you to classify and protect your organization's data, without hindering productivity and collaboration.
4. Sensitivity labels are used to identify the categories of classification types within your organizational data and group the policies that you wish to apply to each category.
5. Build a data map with data lineage processes and insights.
6. Build organizational metadata to support the data assets for business insights and usage.
7. Identify classification and sensitivity labels with Purview's automated scanning.
8. Apply DLP policies against classified and sensitive data to detect risky behavior and prevent unauthorized sharing or breaches.

Next Steps

The completion of this project leaves the Region with new skills and tools to be employed throughout the organization. The following are recommended actions to be taken to further the savings and benefits:

1. Engage additional business units for data governance awareness and communicate the benefits and savings through adoption.
2. Extend the Classification of terms for additional Data Assets, to get more comprehensive classifications.
3. Expand the Business Glossary terms for additional business areas for completeness.
4. Gather Data Lineage for additional Sources and Producers of data, to build a clear picture of all processes.
5. Apply Sensitivity Labeling to additional Data Assets to build a comprehensive data classification.
6. Expand the Data Loss Prevention policies to additional SharePoint sites and One Drive accounts to prove the benefits on data security.
7. Extend the DLP coverage for a selection of Exchange email distribution groups, to measure the potential data loss amounts and determine the impact if expanded across the organization.

Cost Reduction and Efficiencies

The Cost savings have been calculated as part of the engagement, to measure the current and potential cost savings and efficiencies, with detailed calculations included.

1. Savings in the Determination of Classification Terms
2. Savings in the Assignment of Classification Terms
3. Savings on Search for Terms and Classifications
4. Savings on Reduction of Data Losses
5. Savings on Reduction of Data Breaches

Overall Savings and Efficiency Benefits

	Data Catalog In-Scope Proof of Concept	Additional Benefit if Purview Expanded to Rest of Organization
1. Savings in the Determination of Classification Terms	$5320	$152,024
2. Savings in the Assignment of Classification Terms	$10,260	$450,371
3. Savings on Search for Terms and Classifications	$18,275	$346,428
4. Savings on Reduction of Data Losses	$17,400	$326,250
5. Savings on Reduction of Data Breaches	$217,500	$4,078,125
Total Savings	**$268,755**	**$5,353,198**

Appendix

Cost Reduction and Efficiencies (defining classification)

- **Reduction in time for manually determining appropriate classification terms (defining a classification term)**
  - Reduction of time by 95% for Classification Terms
  - The estimate of number of terms has been based on complexity ranking of the business applications
  - The ratio of custom to standard terms is 1:3

Business Area	Business Applications	Classification Terms (Standard)	Classification Terms (Custom)	Average Time to Determine Classification (in hours)	Classification Time (in hours)	Classification Cost Savings (in hours)	Savings
Finance	Water billing data	10	3	0.75	10	9.5	$ 950
Planning	PLANit data	10	3	0.75	10	9.5	$ 950
Transit	Durham 311	10	3	0.75	10	9.5	$ 950
Transit	Spare Labs	7	2	0.75	7	6.65	$ 665
Transit	Trapeze data	7	2	0.75	7	6.65	$ 665
Transit	INIT	7	2	0.75	7	6.65	$ 665
Traffic	REMS data	5	2	0.75	5	4.75	$ 475
Traffic	MATRIX	5	2	0.75	5	4.75	$ 475
Social Services	CASA	7	2	0.75	7	6.65	$ 665
CSIT	GIS	10	3	0.75	10	9.5	$ 950
CSIT	Lansweeper	2	1	0.75	2	1.9	$ 190
Health	Encounter	7	2	0.75	7	6.65	$ 665
	In-Scope Total	**87**	**19**		**56**	**53**	**$ 5,320**
Rest of Organization		1544	515	0.75	1544.25	1467.0375	$ 146,704
		1631	**533**		**1600**	**1520**	**$ 152,024**

Cost Reduction and Efficiencies (assigning classification)

- **Reduction in time for manual assigning classification (assigning values to the Terms)**
  - Reduction of time by 95% for Classifications
  - The Region has 225 databases, with an average of 30 classifiable terms for each application; over 6500 classifications
  - The average time for classification is 45 minutes, taking into account number of staff with input, review period and final decisions

Business Area	Business Applications	Classification Terms (Standard)	Classification Terms (Custom)	Average Time to Determine Classification (in hours)	Classification Time (in hours)	Classification Cost Savings (in hours)	Savings
Finance	Water billing data	30	10	0.75	30	28.5	$ 2,850
Planning	PLANit data	30	10	0.75	30	28.5	$ 2,850
Transit	Durham 311	30	10	0.75	30	28.5	$ 2,850
Transit	Spare Labs	21	7	0.75	21	19.95	$ 1,995
Transit	Trapeze data	21	7	0.75	21	19.95	$ 1,995
Transit	INIT	21	7	0.75	21	19.95	$ 1,995
Traffic	REMS data	15	5	0.75	15	14.25	$ 1,425
Traffic	MATRIX	15	5	0.75	15	14.25	$ 1,425
Social Services	CASA	21	7	0.75	21	19.95	$ 1,995
CSIT	GIS	30	10	0.75	30	28.5	$ 2,850
CSIT	Lansweeper	6	2	0.75	6	5.7	$ 570
Health	Encounter	21	7	0.75	21	19.95	$ 1,995
	In-Scope Total	**261**	**36**		**108**	**103**	**$ 10,260**
Rest of Organization		4633	1544	0.75	4632.75	4401.1125	$ 440,111
		4894	**1580**		**4741**	**4504**	**$ 450,371**

Cost Reduction and Efficiencies (classification Information)

- **Reduction in time to finding Classification information in Purview, as Single Source of Truth**
  - Reduction of time by 85% for locating attribute Classification
  - Reduction of time by 85% for locating attribute Glossary definition
  - Reduction of time by 85% for locating Data Asset information
  - Estimation of Searches per Source is based on frequency of use (Complexity Ranking), with an average 15 minutes per search

Business Area	Business Applications	Complexity Ranking	Number of Searches per year	Average Time to Determine Classification (in hours)	Classification Cost Savings (in hours)	Glossary Cost Savings (in hours)	Data Asset Cost Savings (in hours)	Savings
Finance	Water billing data	10	100	0.25	7	7	7	$ 2,125
Planning	PLANit data	10	100	0.25	7	7	7	$ 2,125
Transit	Durham 311	10	100	0.25	7	7	7	$ 2,125
Transit	Spare Labs	7	70	0.25	5	5	5	$ 1,488
Transit	Trapeze data	7	70	0.25	5	5	5	$ 1,488
Transit	INIT	7	70	0.25	5	5	5	$ 1,488
Traffic	REMS data	5	50	0.25	4	4	4	$ 1,063
Traffic	MATRIX	5	40	0.25	3	3	3	$ 850
Social Services	CASA	7	70	0.25	5	5	5	$ 1,488
CSIT	GIS	10	100	0.25	7	7	7	$ 2,125
CSIT	Lansweeper	2	20	0.25	1	1	1	$ 425
Health	Encounter	7	70	0.25	5	5	5	$ 1,488
	In-Scope Total	**87**	**860**	**95%**	**61**	**61**	**61**	**$ 18,275**
Rest of Organization		1544	15442.5	0.25	1094	1094	1094	$ 328,153
		1631	**16303**		**1155**	**1155**	**1155**	**$ 346,428**

Cost Reduction and Efficiencies (data loss)

- Reduction of data loss through prevention, **average data loss value**
- Estimated cost of **Data Loss** is $20,000 per loss
- Estimated risk of **Data Loss** estimated at 1% per Data Source, with Potential for loss increasing for frequency of access (Complexity Ranking)

Business Area	Business Applications	Complexity Ranking	Potential Data Losses	Risk of Data Loss per Source	Cost of Data Losses
Finance	Water billing data	10	0.1	0.01	$ 2,000
Planning	PLANit data	10	0.1	0.01	$ 2,000
Transit	Durham 311	10	0.1	0.01	$ 2,000
Transit	Spare Labs	7	0.07	0.01	$ 1,400
Transit	Trapeze data	7	0.07	0.01	$ 1,400
Transit	INIT	7	0.07	0.01	$ 1,400
Traffic	REMS data	5	0.05	0.01	$ 1,000
Traffic	MATRIX	5	0.05	0.01	$ 1,000
Social Services	CASA	7	0.07	0.01	$ 1,400
CSIT	GIS	10	0.1	0.01	$ 2,000
CSIT	Lansweeper	2	0.02	0.01	$ 400
Health	Encounter	7	0.07	0.01	$ 1,400
	In-Scope Total	**87**	**0.87**		**$ 17,400**
Rest of Organization		1544	15.4425	0.01	$ 308,850
		1631	**16**		**$ 326,250**

Cost Reduction and Efficiencies (financial loss)

- Reduction of financial and reputational losses through **data breach prevention**
- Estimated cost of **Data Breach** is $5,000,000 per breach
- Estimated risk of **Data Breach** estimated at .005% per Data Source with Potential for loss increasing for frequency of access (Complexity Ranking)

Business Area	Business Applications	Complexity Ranking	Potential Data Breaches	Risk of Data Breach per Source		Cost of Data Breach	
Finance	Water billing data	10	0.005	0.0005		$	25,000
Planning	PLANit data	10	0.005	0.0005		$	25,000
Transit	Durham 311	10	0.005	0.0005		$	25,000
Transit	Spare Labs	7	0.0035	0.0005		$	17,500
Transit	Trapeze data	7	0.0035	0.0005		$	17,500
Transit	INIT	7	0.0035	0.0005		$	17,500
Traffic	REMS data	5	0.0025	0.0005		$	12,500
Traffic	MATRIX	5	0.0025	0.0005		$	12,500
Social Services	CASA	7	0.0035	0.0005		$	17,500
CSIT	GIS	10	0.005	0.0005		$	25,000
CSIT	Lansweeper	2	0.001	0.0005		$	5,000
Health	Encounter	7	0.0035	0.0005		$	17,500
	In-Scope Total	**87**	**0.04**			**$**	**217,500**
Rest of Organization		1544	0.772125	0.0005		$	3,860,625
		1631	**1**			**$**	**4,078,125**

Assumptions in the Calculations

We used the following assumptions in the calculations to assist in matching to industry standards for savings and efficiencies for Purview adoption.

Notes

1 The estimate of number of terms has been based on complexity ranking of the business applications
2 The ratio of custom to standard terms is 1:3
3 The estimate for term definition effort is 45 minutes, or .75 hours per term
4 The estimated cost per hour is calculated from a $800 per day cost
5 The estimated number of information searches is 100 per application per year
6 The estimated number of information searches time savings is 15 minutes per search
7 The potential data losses is a factor of number of applications and employees, with 1% data losses per application, for $20,000 per data loss (Potential data loss source)
8 The potential data breaches is a factor of number of applications and employees, with .05% data breaches per application, with $5Million per breach (Potential data breach source)