부하 Insight。

Outcome Document – Proof of Concept with Microsoft Azure Datalake, Purview and Text Analytics for records disposal

for Te Rua Mahara o te Kāwanatanga Archives New Zealand

30 June 2022

Insight Enterprises NZ Ltd https://nz.insight.com

Document Review

Author	Version	Date	Comment
Richard Malloch	V0.1	02/06/2022	Template with project overview and purpose
JP van Heerden	V1.0	18/06/2022	First draft
JP van Heerden	V2.0	21/06/2022	Indicative costs added
Richard Malloch	V2.1	30/06/2022	Next step recommendation for a production pilot

Reference Documents

Document	Document Name
Archives New Zealand Scope	Data Lakes PoC_Scope_Requirement_Business_Outcome_v1.1.xlsx
Result Sharing Format.xlsx	Excel spreadsheet with output format
MPI metadata - pre-classified and reworked.xlsx	Pre-classified metadata for all documents provided



1 Problem Statement

Archives New Zealand work to ensure effective, trusted government information for the benefit of all New Zealanders. We preserve and protect more than seven million official records, from 19th century treaties to 21st century documents and data. Our goal is for all New Zealanders to easily access and use this taonga, connecting you to your rights and entitlements and stories – now and for the future.

The purpose of this document is to provide the readers with recommendations using the Microsoft Azure platform for the automated disposal, archive or transfer of common corporate records.

The target audience of this document are the project team at the Department of Internal Affairs, including Archives New Zealand, evaluating the suitability of the Microsoft Azure modern cloud platform to automate the disposal process.

For files being retained, there are other added-value use cases provided by the Microsoft Azure modern cloud platform, for example data enrichment using AI powered cognitive services, (an example being video transcription), but these were out of scope for this Proof of Concept.

Problem statement:

Archives New Zealand receive data from New Zealand Government agencies under a General Disposal Authority - developed for the use of public offices wishing to dispose of common corporate public records legally (GDA6) and records of any format that have only short-term transitory value in their immediate and minor facilitation of preparing a more complete public record (GDA7).

Insight were engaged by Microsoft to provide a proof of concept (PoC) evaluating the suitability of the Microsoft Azure modern cloud platform to automate the classification to support the disposal, archive or transfer process.

The key business outcomes to be proven during this proof of concept were:

- Ensuring the correct policy for the retention and disposal of all files determined automatically (for GDA 6 and GDA 7)
- Information that may be of interest to Māori is identified
- Security classification (PSR) that are automatically applied to the provided data is accurate
- Manual intervention and input into business processes (DA) minimised

Sample data from MPI was used for in this proof of concept, with testing provided by the project team at Archives New Zealand / Department of Internal Affairs. Insight acknowledges the valuable support and guidance from this team.

2 PoC Solution & Products Used

This section provides an overview of the proposed solution architecture used for this proof of concept evaluating the suitability of the Microsoft Azure modern cloud platform to automate the disposal process.

Please note that the solution architecture below is not intended to be a complete solution, but rather a demonstration (in a limited amount of time) of the capabilities of the products involved.

If a Microsoft platform is selected for a final solution, design of a more detailed architecture will be required.



2.1 PoC Architecture

Figure 1: High-level Solution Architecture for Archives New Zealand PoC Platform



The following table summarises the solution components and associated toolsets:

Component	Description	PoC Toolsets
Collaboration Platform	 SharePoint is Microsoft's collaboration portal that facilitates documents storage and management, allows teams to share knowledge, quickly find information and empowers teamwork. Microsoft Purview Compliance Portal provides easy access to the data and tools you need to manage to your organization's compliance needs. Amongst other functionality, the Compliance Portal allows users to capture and manage Information Protection labels. The portal can also be used to configure automatic tagging of documents with Information Protection labels. Microsoft Office applications allow users to capture Managed Information Protection (MIP) labels against documents stored in SharePoint or other locations. 	 Microsoft SharePoint Microsoft Purview Compliance Portal Microsoft Office applications
Document transfer mechanism	A mechanism may be required to copy or transfer documents from SharePoint to Azure Data Lake. For this PoC, the documents were copied manually as this process was out of scope. (This can be automated).	• Manual copy
Data Storage	An Azure storage account allows your organisation to store a variety of object, including blobs, file shares, queues, tables, and disks and provides a unique namespace for your Azure Storage data that's accessible from anywhere in the world. A storage account can be configured to act as a Data Lake, allowing you to store many documents, with built in redundancy. Storage accounts act as a source for other components, such as Purview and Text Analytics.	 Azure Storage Account (to provide the datalake)

Component	Description	PoC Toolsets
Data Governance	 Microsoft Purview Data Governance Portal provides a unified data governance service that helps you manage your on-premises, multi-cloud, and software-as-a-service (SaaS) data. The Microsoft Purview governance portal allows you to: Create a holistic, up-to-date map of your data landscape with automated data discovery, sensitive data classification, and end-to-end data lineage. Enable data curators to manage and secure your data estate. Empower data consumers to find valuable, trustworthy data 	• Microsoft Purview
Document Classification	Azure Text Analytics is a collection of features from Cognitive Service for Language that extract, classify, and understand text within documents.	Azure Text Analytics
Integration Application	In the above solution, Text Analytics has been used to classify documents. These documents would already have been scanned by Purview and a mechanism would update Purview with the additional classifications.	Manual capture
Extraction Application	Purview provides a fully functional portal to browse and search for documents, but a mechanism to extract the document lists and classifications may be required.	API extraction

Table 1: PoC Solution Components

Further information:

• Azure data services (data lake and data ingestion),

https://azure.microsoft.com/en-us/solutions/data-lake/

• Microsoft Purview data catalog (data classification),

https://azure.microsoft.com/en-us/services/purview/#overview

• Microsoft Azure Text Analytics service,

https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/#overview

• Microsoft Purview Trainable Classifiers for Managed Information Protection (MIP)

 $\underline{https://docs.microsoft.com/en-us/microsoft-365/compliance/classifier-learn-about?view=o365-worldwide}$



3 Application of the products used

3.1 SharePoint & Information Protection Labels

The Microsoft Purview Compliance portal can be used to capture and add Managed Information Protection (MIP) labels. These labels can then be applied to documents in one of two ways:

- Automatically through several data classification mechanisms, such as Trainable Classifiers.
- Manually by capturing the labels against Office documents

The scope and timing of this PoC did not allow for extensive demonstration of the above automated capabilities, but this offering is a core part of Microsoft's governance offering.

It should be noted that no sensitive documents were used for this PoC and that MIP labels were applied for demonstration purposes only.

	Microsoft Purview									? ھ	ЭН
=											
٩	Permissions		Infor	nation protection						😒 Remove from navigation	
咨	Trials		Overview	Labels Label policies Au	ito-labeling						
Sol	utions										
₽	Catalog		U too can now create sensitivity abels was privacy and access control settings for learns, Shareholm state, and Microsoft 305 Groups. To do this, you must that complete these steps to enable the teature. Sensitivity labels are used to classify email messages, documents, sites, and more. When a label is applied (automatically or by the user), the content or site is protected based on the settings you choose. For example, you can create labels that encrypt								t
R	Audit		+ Create	a label 🖵 Publish label 🖒 Refres	h	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,				11 items	
Q	Content search			Name		Order	France	Created by	Last modified		
G	Communication compliance		-	Percond		0 - lowert	Scope	created by	Last modified		
G	Data loss prevention			Personal		0 - Iowest	File, Email		may 22, 2022 605:59 HM		
Ŕ	eDiscovery	-		Public		1	File, Email		May 22, 2022 8:05:59 AM		
8	Data lifecycle management		$\Box \rightarrow$	General		2	File, Email		May 22, 2022 8:05:59 AM		
G	Information protection		$\Box \rightarrow$	Confidential		5	File, Email		May 22, 2022 8:06:01 AM		
G	Information barriers	/	$\Box \rightarrow$	Highly Confidential		9	File, Email		May 22, 2022 8:06:07 AM		
₽ <u></u>	Insider risk management			In Confidence		12	File, Email, Schematized data assets	JP Van Heerden	Jun 13, 2022 8:36:42 AM		
R	Records management			Restricted		13	File, Email, Schematized data assets	JP Van Heerden	Jun 13, 2022 8:39:16 AM		
8	Priva Privacy Risk Managem 🚿	/		Sensitive		14	File, Email, Schematized data assets	JP Van Heerden	Jun 13, 2022 8:39:53 AM		
Ea	Priva Subject Rights Requests			Secret		15	File, Email, Schematized data assets	JP Van Heerden	Jun 13, 2022 8:40:21 AM		
۲	Settings			Top Secret		16	File, Email, Schematized data assets	JP Van Heerden	Jun 13, 2022 8:40:39 AM		0
()	More resources			Restricted2		17 - highest	File, Email, Schematized data assets	JP Van Heerden	Jun 13, 2022 8:12:21 PM		ç.
0											

Figure 2: Information Protection labels in Purview Compliance portal

	SharePoint	P Search	S (1)
6 #	AP Archives-PC		Private group ★ Following 🕺 1 member
	Home	🕂 New 🗸 🔻 Upload 🗸 🧮 Edit in grid view 😰 Share 🕫 Copy link 🤤 Sync 🛓 Download 🖏 Add shortcut to OneDrive 🖪 Export 🗸 🤗 Automate 🗸 \cdots	= All Documents 🗸 🛛 🗸
•	Conversations		
U	Documents	Documents > Security	
	Shared with us	□ Name ∨ Modified ∨ Modified By ∨ + Add column ∨	
	Notebook	Demo - ET324380_18 10 2006 - Felix Donnelly College.DOCX Sunday at 101 PM JP Van Heerden	
	Pages	Demo - Experimental use.5july2006.doc Sunday at 101 PM JP Van Heerden	
	Site contents	Demo - import restrictions.20Dec2006.doc Sunday at 1.01 PM JP Van Hearden	
	Recycle bin	Demo - JT.FlatWhite.1July2006.doc Sunday at 1.01 PM JP Van Heerden	
	Edit	Demo - Specific topics for Hon Ruth Dysons speech.dock Sunday at 101 PM JP Van Hearden	
		Demo - Talking points for Minister's photo opp 26 Oct.doc Sunday at 101 PM JP Van Heerden	
		ET324380_18 10 2006 - Felix Donnelly College.DOCX Sunday at 1233 PM JP Van Hearden	
		Experimental use-5july2006.doc Sunday at 1233 PM JP Van Hearden	
		import restrictions.20Dec2006.doc Sunday at 1233 PM JP Van Hearden	
		JT.FlatWhite.1July2006.doc Sunday at 1233 PM JP Van Heerden	
		Specific topics for Hon Ruth Dysons speech.docx Sunday at 12:33 PM JP Van Heerden	
		Talking points for Minister's photo opp 26 Oct.doc Sunday at 12:33 PM JP Van Heerden	



Figure 3: SharePoint documents



Figure 4: Manually capture Information Protection labels

3.2 Transferring data to the Data Lake

A mechanism can be applied to copy or transfer documents from SharePoint to Azure Data Lake. For this PoC, the documents were copied manually as this process was out of scope

3.3 Scanning Documents with Purview

Once document have been copied to the Data Lake, Microsoft Purview can be used to scan the documents. Microsoft Purview can be configured to pick up Information Protection labels that have been applied to documents and to apply classification rules (such as iwi and hapū names) that been configured in Purview.

Micr	osoft Purview Governance Portal + Arch	ves PoC . P Search catalog	<⊐ 123 Q ⊚ ? R jøvanheerden@tsjöw.co 🗩
>>	Data catalog >		
н.	Browse assets		
	🕐 Refresh		
	By collection By source type		
۷	< View collection tree	Source type : all) Instance : all) 😨 Clear all filters	
	Data Lake Archives PoC >	Showing 1-2 out of 2 results	Sort by: Relevance
	Sub collection(s)	Manual - JT.FlatWhite.1July2006.docx (Top Secret)	
	Related Archives PoC (Parent)	Acure Data Lake Storage Gen2 Path Https://archivesauepocadio31.dfs.core.windows.net/files/Manual/Manual - JT.FatVMnite.1July2006.docx	Updated 3 days ago
	Narrow results by:	Manual - Talking points for Minister's photo opp 26 Oct.docx (Top Secret)	
	* Object Type	Azure Data Lake Storage Gen2 Path	Under all designs
	OD Data pipelines Piles Folders OD Giosary terms Alt Reports		
	End Stored procedures End Tables		
	* Classification		
	Secret Ø		
	* Contact		
	* Label ····		
	Top Secret		
	in Confidence		
	Restricted2		

‡ Insight

Browse assets 🕐 Refresh ¢ By collection By source type ۲ Source type : all Instance : all 🛛 😨 Clear all filte 0 Data Lake Showing 1-1 out of 1 results Sort by: Relevant Ngati Rangiteaorere 19 Azure Data Lake Storage Ger Concessegocadis01.07 Sub collection(s) Related Archives PoC (Parent Narrow results by: * Object Type ObjectType Object Type Object Type Object Type Object Typ Secret Restricted Confidential Sensitive Top Secret 0

Figure 5: Purview assets filtered by Information Protection labels

Figure 6: Purview assets filtered by classifications for a specific iwi

3.4 Classifying documents with Text Analytics

Text Analytics can be used to create and train a model so that it can predict document attributes such as Disposal Authority, Class and Action. The larger and more diverse the set of documents used to train the model, the more accurate the predication. The 1,000 documents used in this PoC provided a high degree of accuracy.

Cognitive Services Languag	je Studio		Q 😳 🏟	? jpvanheerden@tsjbw.onmicrosoft.com archives-aue-lang3-dev (australiaeast, S)				
*	Language Studio > Custom Text Classification > Archive_Class3 - Model	performance > Archive_Class_Model3						
🟠 Language Studio	Archive_Class_Model3							
Projects	Classes that are neither labeled nor predicted in the test set will not be part of the displayed results. Learn more about data splitting,							
Archive_Class3	Overview Test set confusion matrix							
Data labeling	Status: 🖉 Trained successfully							
Training jobs	Finished training on: 6/16/2022, 3:24:53 PM	Precision	Recall	F1 score				
Model performance	Total training time: 3 hour(s), 42 minute(s), 0 second(s)							
🍄 Improve model		87.4%	87.4%	87.4%				
Deploying a model								
Testing deployments	Class performance metrics							
Project settings	Class name \vee	Precision % \smallsetminus	Recall % \smallsetminus	F1 score \vee				
	1.3.1	88.64	88.64	0.8863635659217834				
	1.1.2	83.75	90.54	0.8701297640800476				
	1.2.5	71.43	50.00	0.5882352590560913				
	4.1.1	94.12	94.12	0.9411764740943909				

Figure 7: Text Analytics model

3.5 Enriching Purview Assets

A mechanism is required to enrich the list of documents catalogued by Purview. During this PoC, the documents were manually updated. Azure Text Analytics provides an API interface where a document can be submitted and an attribute returned. An application can be developed to do this and to update Purview, potentially using the Business Glossary.



Cognitive Services	anguage Studio 🛛 🕞 🛞 ? (invanheurden@typesonmicrosoff.com g
	« Language Studio > Custom Text Classification > Archive_Class3 - Testing deployments
放 Language Studio	Testing deployments
Projects	Test a deployment by providing sample text to find out what class it gets categorized as. These are the same predictions as when making API calls against your model in code. Learn about the info in these API requests and responses,
Archive_Class3	A Run the test
Data labeling	Deployment name
🚳 Training jobs	Accuracy several s
Model performance	Enter your own text, or upload a text document
Improve model	Ministry of Agriculture and Forestry
Deploying a model	MAE Critical Process Owner Sign-off (#1) of Year 2000 Business Continuity Plans
A Testing deployments	Further to my responsibility as Business Process Owner for the following critical business process,
Project settings	I hereby provide assurance to the Director General. Professor Bruce Ross, that:
	2960/125000 Drag and drop. htt file(s) here or Browse for a file
	Result JSON
	11: Sort \sim
	1.2.1 Confidence 51.00%



Micro	soft Purview Governance Po	ortal 🔸 Archives PoC			\$ 1	a 🧔	© 1	~	jpvanheerden@tsjbw.on TSJBW
>>	Data catalog > Browse assets >	Azure Data Lake Storage Gen2 > a	chivesauepocadls01 > 55.doc >						
•	Edit "55.doc"								
	Overview Schema Con	tacts							
•	Name *	55.doc							
	Asset description								
				h.					
	Classifications ①	Select		\checkmark					
	Glossary terms ①	Class 1 2 1		\sim					
	Certified 🕕	Class 1 1 2							
		Class 1 2 1							
	Save Cancel								

Figure 9: Updating Purview manually

3.6 Extracting data from Purview using APIs

While Purview provides a fully functional portal to browse and search for documents, a mechanism to extract the document lists and classifications may be required. Developing this mechanism is out of scope for this PoC, but Purview does provide a set of APIs that can be used to extract a list of documents with their labels, glossary terms and classifications.



Postman File Edit View Help					- 6	1 ×
🥖 Home 🛛 Workspaces 🗸 🛛 API N	Network V Reports	Explore	Q Search Postman	🖧 🦂 Invite	හී 🔅 🗘 🧿 Upgra	de v
A My Workspace	New Import POST Get	Token POST List Assets	+ 000		No Environment	~ ©
Collections + =	°°° Purvie	w / List Assets			🖺 Save 🗸 👓 🥖 🗐	Þ
Purview Post Get Token	POST	v https://archives-aue-dev-pview-01	.purview.azure.com/catalog/api/search	query?api-version=2022-03-01-preview	Send	Ē
APIs POST List Assets	Params	s Authorization Headers (9) Body	Pre-request Script Tests	Settings	Cookie	s
Environments	Туре	0Auth 2.0 ~	Heads up! These parameters hold sen	sitive data. To keep this data secure while working in re about variables A	a collaborative environment, we $\qquad \times$	(j)
Canal Mock Servers	The auti generate Learn m	horization data will be automatically ed when you send the request. nore about authorization a This	ent Token access token is only available to you. Sync	the token to let collaborators on this request use it.		: ():
а _t ⁰ Flows Ф History	Body C Pretty 1 2 3 3 4 5 6 7 7 8 9 9 10 11 12 13 14 15 16	Cookies Headers (12) Test Results / Raw Preview Visualize JSON * "value": [* "value": [* "omner": "\$superuser", * "qualifiedKame": "https:// * "entityType": "azure_data * "https:// * "classification": [* "Secret"], * "classification": [* "Secret"], * "azure Data Lake Store	A v T Archivesauepocadis01.dfs.core. Lake_gen2_path*, 180, Assc_cibE_48c0_9050_AE957A4A366 age Gen2*	Status: 200 OK Time: I windows.net/files/Manual/Manual - JT.FlatW a*	307 ms Size: 5.8 KB Save Response	

Figure 10: API to extract Purview assets



4 Model Used and Results

Sample documents were used to train the text analytics model. There was a big difference in the accuracy of the model between the 100 documents used for the initial testing and the full set of 1,400 sample documents provided.

Training for classification fields using Purview – example Disposal Authority

Number of Files Used	Time to model	train	Model accuracy
100	1 hour		35-40%
1400	3 hours		75-85%

It is expected that the accuracy would increase with a larger set of documents to train the model.

The classifiers used to match a field against field in a reference list, for example the name of an iwi, took about 5 minute each to train.

4.1 Data deleted

The sample data provided for the PoC has been deleted:





MPI data for Data Lakes project.zip Done



OneDrive_2022-06-08.zip Done



5 Road Map

Given the success of the proof of concept, we are recommending the following options are the next step. We recommend option 1, being a production pilot.

Option 1: Production Pilot with MVP use case(s)

We recommend a production pilot as the next step using one government agency partner and defined use cases for GDA 6 and GDA 7.

Once that is proven, you will be ready to move to a full production rollout – adding additional use cases, or additional agencies.

Stages:

- a) Discovery:
- For the MVP use case(s), to validate the business process for each agency involved, roles and responsibilities
- architecture validation
- design of integration to external systems
- design of disposal method
- b) Foundations
- Establish the technology foundation for dev, test and prod
- c) MVP use case(s)
- Configure, test and deploy MVP use case(s)

Investment: will vary depending on the use case(s) selected.

Option 2: Proof of concept for additional use case(s)

This PoC evaluated primarily the use cases for GDA 6 and GDA 7, and lightly for other classifications. You have the option to invest in additional uses cases (for example MPI ontology) before moving to a production pilot.

It would require uses cases more fully documented to estimate the investment for each use case.



6 Indicative Cloud Pricing

The pricing below relates to the estimated volumes and functionality provided by the PoC. The pricing only covers Azure costs as we have established that the agencies involved are likely to have Microsoft 365 E5 licenses, which cover SharePoint and other Microsoft 365 information protection functionality.

Service type	Custom name	Region	Description	Estimated monthly cost
Storage Accounts	Storage, hot tier, 10TB	Australia East	Block Blob Storage, General Purpose V2, LRS Redundancy, Hot Access Tier, 10 TB Capacity - Pay as you go, 10 x 10,000 Write operations, 10 x 10,000 List and Create Container Operations, 10 x 10,000 Read operations, 100,000 Archive High Priority Read, 1 x 10,000 Other operations. 1,000 GB Data Retrieval, 1,000 GB Archive High Priority Retrieval, 1,000 GB Data Write	\$315.24
Microsoft Purview	Daily scan < 1h/day ADLS; 1 CU;	Australia East	Elastic Data Map: 1 Capacity Unit hour, 720 hours, Automated Scanning and Classification: 1 Total scan duration in hour x 32 Total vCores across scans (For other data sources), Other features: 0 Resources Set hours, Microsoft Purview Data Catalog: C1 Service	\$483.82
Azure Cognitive Services	Custom TA - Classification API; 10k rec/day; <10h training/month; 2 endpoints	Australia East	Cognitive Services for Language, Pay as you go, Standard, Cognitive Services for Language: 30 x 1,000 text records, Text Analytics for health: 0 x 1,000 text records, Custom question answering and prebuilt question answering: 0 x 1,000 text records, Custom named entity recognition (preview) and Custom text classification (preview): 10 x 1,000 text records, 10 training hours, 2 endpoint hosting models	\$169.91
Support			Support	\$0.00
			Licensing Program	Microsoft Customer Agreement (MCA)
				<u>toco 07</u>
			IOTAI	\$968.97

Table 2: Costs

Detailed costs estimates are difficult to provide, given that they will be dependent on volumes. It is proposed that cost estimates are calculated as part of the next production pilot project.

\$\$ Insight.