Data extraction & processing while maintaining security, monitoring & standard on azure platform.

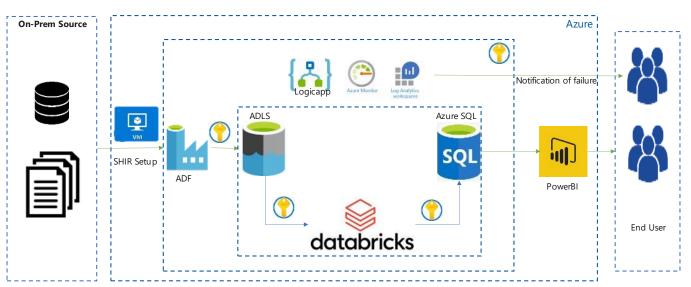
Problem statements: As data engineer usually will get different requirement for data processing & transformation of various sources, Azure offers a wide range of services and tools to help you achieve this.

We received requirement to pick up files from on premises to azure cloud platform & perform checks on data quality while maintaining security, monitoring & data lake transformation layers before loading data in Data warehouse.

Services use:

- 1. Azure account: For using cloud platform
- 2. Azure Data lake : For storing data & maintaining transformation layer
- 3. Azure DataFactory:- For data ingestion from onprem to azure & data processing
- 4. Azure Databricks: For data transformation & delta tables
- 5. Azure Synapse: For DWH & data modelling purpose
- 6. Azure Keyvualt: For maintaining credentials & secrets at centralize location.
- 7. Azure monitor : For monitor Azure resources
- 8. Azure log analytics: For querying logs
- 9. Logic Apps: For sending email on failure & success
- 10. Azure VM : For setup SHIR

Architecture Diagram:



Logical architecture

Technical Requirement:

Setup azure environment with listed services

•	Azure Data lake setup & maintain 3 different layer in ADLS					
	RAW: Stora raw data from source					
	Refined: store cleansed data after processing					
	Processed : store transformed data					
	Carting A characteristic b Datas antista					
	+ Container 🔒 Change access level 🤌 Restore containers					
	Search containers by prefix					
	Name					
	¢lara					
	\$logs					
	proccesssed					
	raw					
	refined					

2. Setup vm for selfhosted IR to connect on prem server to pickup file

Name ↑↓	Type ↑↓	Sub-type \uparrow_{\downarrow}	Status ↑↓
👆 AutoResolveIntegrationR	Azure	Public	🗸 Running
integrationRuntime1	Self-Hosted		🗸 Running

3. <u>Setup Azure data factory</u> for data ingestion & proccesessing :

		ForEach	אל	2
Get Metadata	~	ForEach_1		Notebook
Get Metadata_1		Activities	Ø	Notebook1
		$ \begin{array}{c} (\chi) \\ \text{Set} \\ \text{variable1} \end{array} \longrightarrow \begin{array}{c} \textcircled{1} \\ \text{Get} \\ \text{Metadata2} \end{array} \longrightarrow \begin{array}{c} \cancel{1} \\ \text{If} \\ \text{Condition1} \end{array} \longrightarrow \textcircled{+} \end{array} $		

Below checks has been performed in ADF for data consistency:

• Check if the file is available in the path. If it's not available, there should be timeout after 1 minute:

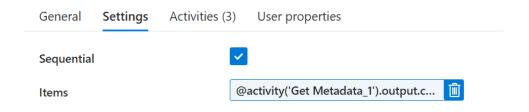
General Settings User pro	perties	
Name *	Get Metadata_1	Learn more 🖸
Description	fetch from blob storage	
Activity state (preview) $^{(\mathrm{i})}$	Active Inactive	
Timeout ⁽ⁱ⁾	0.00:01:00	
Retry ^①	10	

• Check if the file size is greater than 20b or not. If the file size is greater than 20b then needs to process:

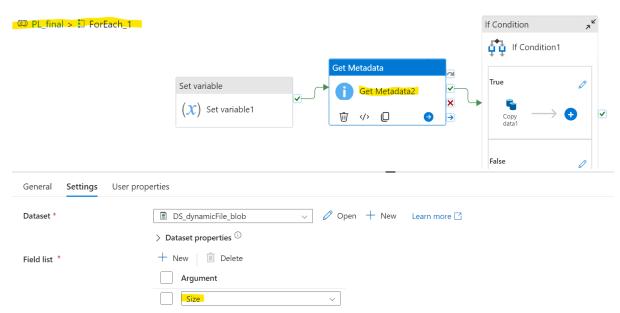
We have used 2 get metadata to check first list of files & then stored file name in variable to check size of each file

(ForEach	
Get Metadata Get Metadata_1	ForEach_1	
ŵ □ 🥯	Activities	v
	$ \begin{array}{c} (\chi) \\ \text{Set} \\ \text{variable1} \end{array} \longrightarrow \begin{array}{c} \textcircled{\bullet} \\ \text{Get} \\ \text{Metadata2} \end{array} \longrightarrow \begin{array}{c} \swarrow \\ \text{If} \\ \text{Condition1} \end{array} \longrightarrow \begin{array}{c} \textcircled{\bullet} \end{array} $	
General Settings User prop	erties	
Dataset *	🗈 DS_capstone_blob_getmatadata1 🗸 🖉 Open 🕂 New Learn more 🛛	
Field list *	+ New Delete	
	Argument	
	Child items ~	

Get metadata output passed to foreach loop activity to run loop for every file



Inside for each loop used variable to store file name & with help of 2nd getmetadata pick size of file.

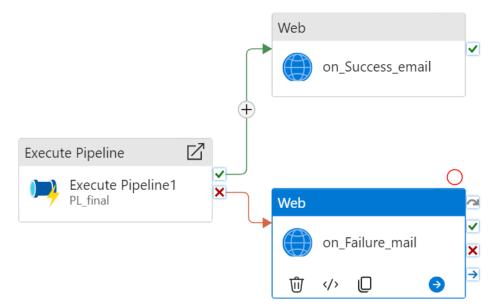


Passed size condition in ifcondition activity to pick files which are greater then 20b

@greaterOrEqua	ls(activity('Get Metadata2').output.size,21)	
ගා PL_final > :] ForEach_	1	If Condition1
	Set variable Get Metadata (\mathcal{X}) Set variable1 Get Metadata2	$ \begin{array}{c} \hline Copy \\ data \end{array} \longrightarrow \begin{array}{c} \bullet \\ \hline \end{array} \end{array} $
General Activities (1)	User properties	
Expression ⁽ⁱ⁾	@greaterOrEquals(activity('Get Meta	
Case	Activity	
True	Copy data1	

• Setup email configuration on failure & success of pipeline to get notify

Create master pipeline to call logic app to get notified on pipeline failure



4. **<u>Configured Logicapp to trigger email</u>** & passed logicapp url in be activity:

Validate Debu	g 续 Add trigger Web ① 小 D	<pre>{ "message" : "Pipeline with run ID @{pipeline().RunId} is successfully executed.", "dataFactoryName" : "@{pipeline().DataFactory}", "pipelineName" : "@{pipeline().Pipeline}", "receiver" : "@{pipeline().parameters.receiver}" }</pre>
General Settings	Execute Pipeline	Clear contents Activity outputs Parameters System variables Functions Variables P Search
URL * ⁽) Method * ⁽) Body	Https://prod-01.westus2-logic azure com:l. A Information will be sent to the URL specified. Please ensure you trust the URL entered. POST Yors ("message" : "Pipeline with run ID @	Execute Pipeline1 Execute Pipeline1 activity output Execute Pipeline1 Execute Pipeline1 pipeline return value
Authentication ${}^{\bigcirc}$	None ~	OK Cancel

5. Setup keyvault to store credentials

Kvnt Secrets ☆				
	« 🕂 Generate/In	nport 💍 Refresh Res	tore Backup View sampl	e code 🛛 😶
🗳 Tags	*			
🔀 Diagnose and solve problems	Name	Туре	Status	E
≸∃ Access policies	adbtoken		✓ Enabled	
🗲 Events				
Objects				
🕈 Keys				
Secrets				
Certificates				

count selection method ① From Azure subscription ④ Enter manually RL *	tnentication type	
IRL * https://chaitengoodp.sconceance_fs.core.windows.net/ Storage account key Azure Key Vault AKV linked service * LS_kv_copatione Secret name * Copatione - datalake G Edit	ccount key	
IRL * https://chaiteny.oucpuscenceance_fs.core.windows.net/ Storage account key Azure Key Vault AKV linked service * LS_kv_copatione Secret name * Copatione - datalake GE Edit	count selection method	0
https://chaitergourgenenerates=lis.core.windows.net/ Storage account key Azure Key Vault AKV linked service * LS_kv_copatione Secret name * Copatione - datalake C Edit) From Azure subscription	 Enter manually
Storage account key Azure Key Vault AKV linked service * ① LS_kv_expetime Secret name * ① oxpetime-datalake ✓ Edit	JRL *	
AKV linked service * 0 LS_kv_copstone Secret name * 0 copstone-datalake ✓ Edit	https://chaitanyacapacco	ierake 1/5.core.windows.net/
AKV linked service * 0 LS_kv_copstone Secret name * 0 copstone-datalake ✓ Edit		
LS_kv_expatione Secret name * ① copatione-datalake		
Secret name * ① copystone-datalake Copystone-datalake Copystone-datalake	Storage account key	Azure Key Vault
⊂e pstone -datalake ✓ Edit		Azure Key Vault
✓ Edit	AKV linked service * ①	Azure Key Vault
-	AKV linked service * ⁽¹⁾ L5_kv_ capstone	Azure Key Vault
Secret version ①	AKV linked service * ① LS_kv_expstone Secret name * ①	Azure Key Vault
Secret version 🛈	AKV linked service * ① LS_kv_expectance Secret name * ① (**********************************	Azure Key Vault
	AKV linked service * ① LS_kv_expotence Secret name * ① Compatione-datalake Compatione-datalake Compatione-datalake	Azure Key Vault

Once all checks has performed databricks notebook will be executed from datafactory pipeline to maintain transformation layer in adls

Get Metadata Get Metadata_1	ForEach ForEach_1 Activities (χ) Set Variable1 Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get Get G	x ⁴	Notebook Notebook1	
General Azure Databricks Setti Databricks linked service *		Edit + New		

Setup Azure databricks :

in Databricks couple of transformation applied , I am attaching some of them only as per confidentiality,

- If date is NULL or blank, give default date as '2020-11-28'. Format of date column should be YYYY-MM-DD.
- Remove the entries which has URL field value as 'ERROR'.
- Transform the values of column country with their acronyms. For eg: Austria would be replaced by 'AUST', Belgium by 'BELG' etc. Y

```
Python >
```

```
1 #read the parquet file from adls
2 source_file="abfss://refined@chaitanyacapastonelake.dfs.core.windows.net"
3 df_source=spark.read\
4         .format('parquet')\
5         .option('inferschema',True)\
6         .load(source_file)
7
8 adf=df_source
```

Cmd 3

1 # Apply transformations
2

```
3 target_df = adf.withColumn("date", when(col("date").isNull(), "2020-11-28").otherwise(col("date")))
4 target_df = target_df_filter(col("url") '= "FRROR")
```

```
4 target_df = target_df.filter(col("url") != "ERROR")
5
```

```
24 #run the loop
    for rows in table1.find_all('tr'):
25
26
      column=rows.find_all('td')
27
     if len(column)>=2:
28
       country_name=column[0].text.strip()
29
        acronym=column[1].text.strip()
30
        country_acronym[country_name]=acronym
31
    #covert the column 'country' of transformed table to upper-case
32
33
     target_adf=target_df.withColumn("country",upper(col("country")))
34
35
    # Create a DataFrame from the acronym data
     acronym_df = spark.createDataFrame(country_acronym.items(), ["country", "acronym"])
36
37
38
    # Join the original DataFrame with the acronym DataFrame to replace values
    adf1 = target_adf.join(acronym_df, "country", "left").select("*")
39
40
    # Interchange the positions of two columns (e.g., swap "Age" and "Country") and then drop country table
41
42
    result_df = adf1.withColumnRenamed("country", "temp").withColumnRenamed("acronym", "country").withColumnRenamed
     ("temp", "acronym").drop("acronym")
43
```

Complete code & ARM template added in github.

Once data transformation done loading data in sql to build pbi dashboard on so can be consumed be end user.

Challenges in implementing solution:

integrating all azure services together in secure way with help of vnet & keyvault so data can be processed from on prem to azure & load successfully in database to consume.

In while process criticality of data is key things to protect so overall flow will work smoothly as per architecture diagram.

Business Benefits:

In all project as data engineering will play a key role to perfoem various data operation & maintain security so its help of this high level flow can implement security best practices can be followed.

Github link:

https://github.com/deeksharm/azuredataengineer