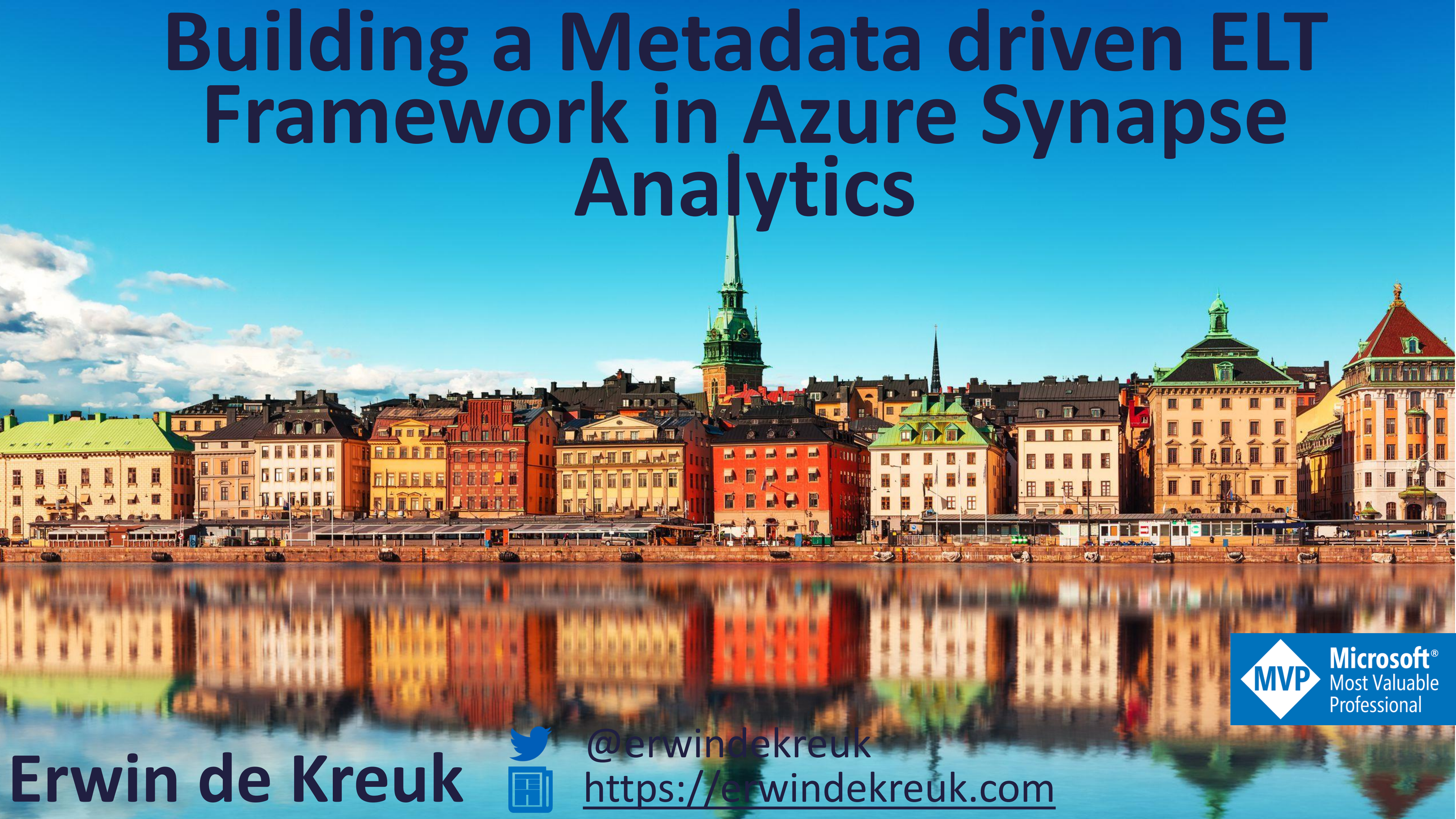
Building a Metadata driven ELT Framework in Azure Synapse Analytics



Erwin de Kreuk

@erwindekreuk

https://erwindekreuk.com

Microsoft® Most Valuable Professional

Erwin de Kreuk

Principal Consultant | Lead Data and AI | Data Platform MVP | Public Speaker …

Erwin de Kreuk

Principal Consultant – Lead Data & AI
InSpark

@erwindekreuk

linkedin.com/in/erwindekreuk

erwindekreuk.com github.com/edkreuk

Microsoft® Most Valuable Professional

b3

REBTECH

SQLSERVICE

24

solarwinds

dbWatch

redgate

devart

Snabbfoting

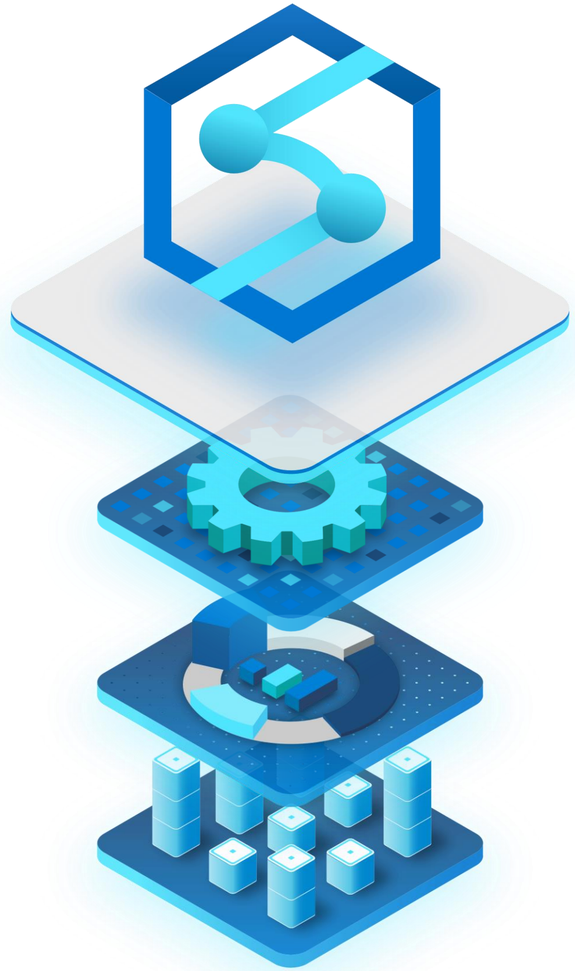
INSPARK

We Are InSpark

We help organizations **accelerating** their **digital transformation** with impactful Microsoft solutions & expertise

INSPARK

Azure Synapse Analytics



Metadata driven ELT Framework

Objectives of today

- Why
- Out-of-the-box Framework
- Custom-made Framework
- Recap

Why

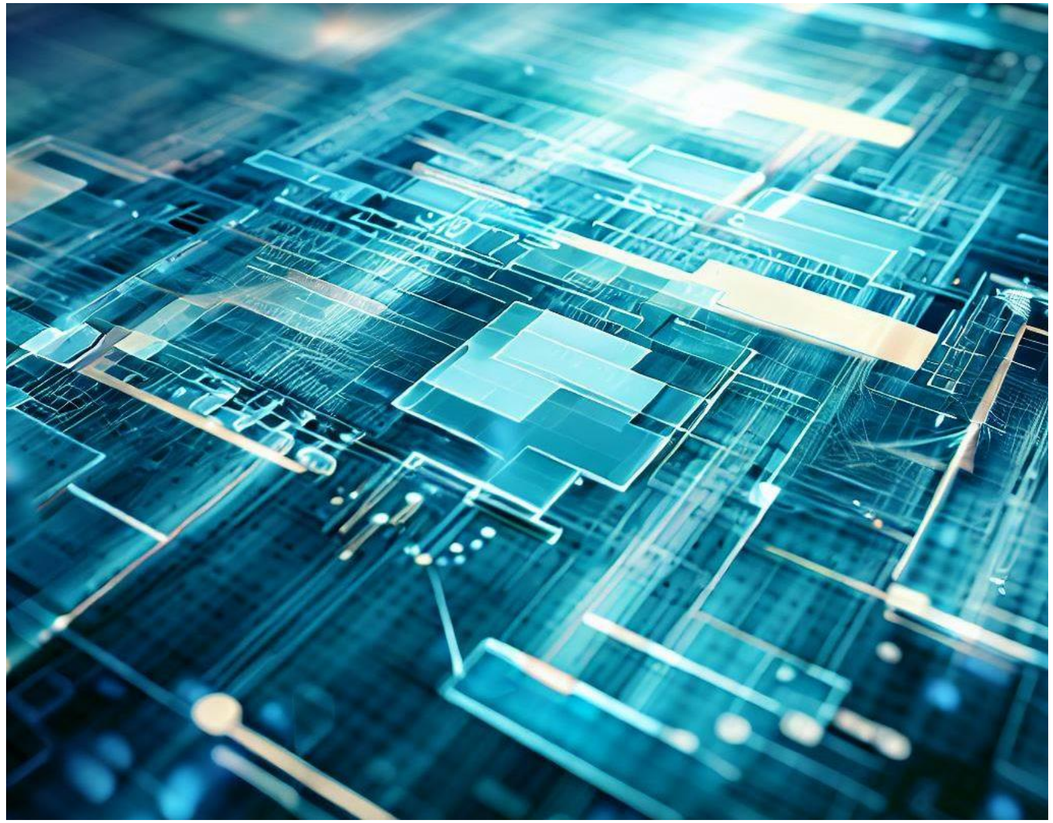
- **Centralized** metadata management for better governance and maintenance.
- **Flexibility** and **agility** by separating metadata from code.
- **Scalability** and performance leveraging Synapse's distributed processing capabilities.
- **Reusability** and standardization of data integration components.
- **Simplified** maintenance through metadata modifications instead of manual code changes.
- **Collaboration** and documentation for alignment between business and technical teams.
- **Extensibility** and customization using Synapse's integration capabilities.

INSPARK

Out-of-the-Box Framework

Out-of-the-Box

- Ready-to-use.
- Rapid implementation.
- Limited customization.
- Lower development effort.
- Lower upfront costs.
- Ongoing support and updates.



Custom-Made Framework

Custom-Made

- Tailored to specific needs.
- Full control over design and features.
- Higher development effort.
- Flexibility and extensibility.
- Higher upfront costs.

INSPARK

Out-of-the-Box Framework

- Available by default in Azure Synapse Analytics and Azure Data Factory
- Azure SQL Database as requirement (Linked Service)
- Almost next, next, finish

Metadata-driven copy task

You will get parameterized pipelines which can read metadata from an external store to load data at a large scale.

INSPARK

Metadata-driven Framework

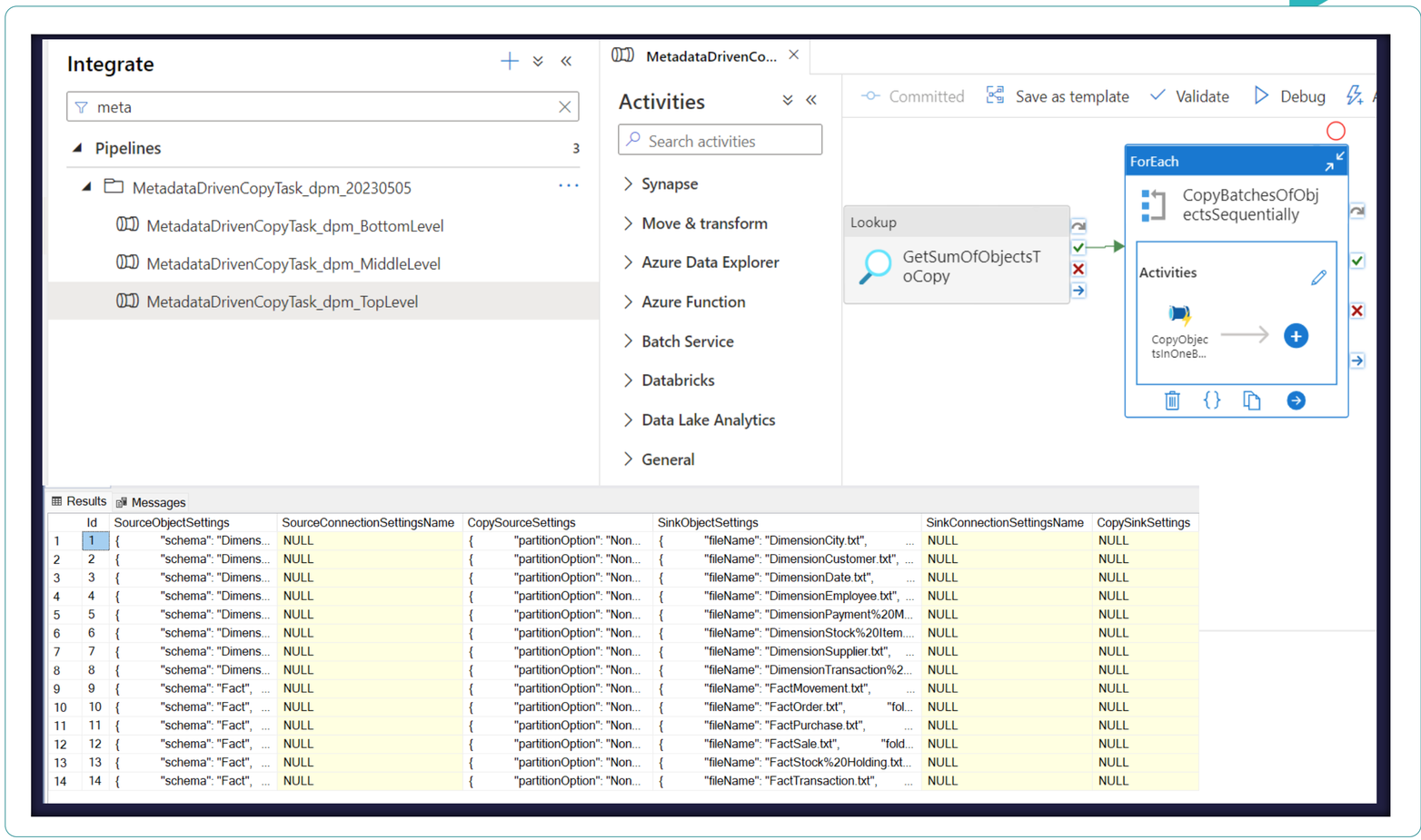
Out-of-the-Box Framework

- Connect to table data store
- Define Schema and Table name
- Define Schedule
- Connect to Source
- Select Tables
- Define full or Incremental
- Define Destination
- Create tables in datastore

INSPARK

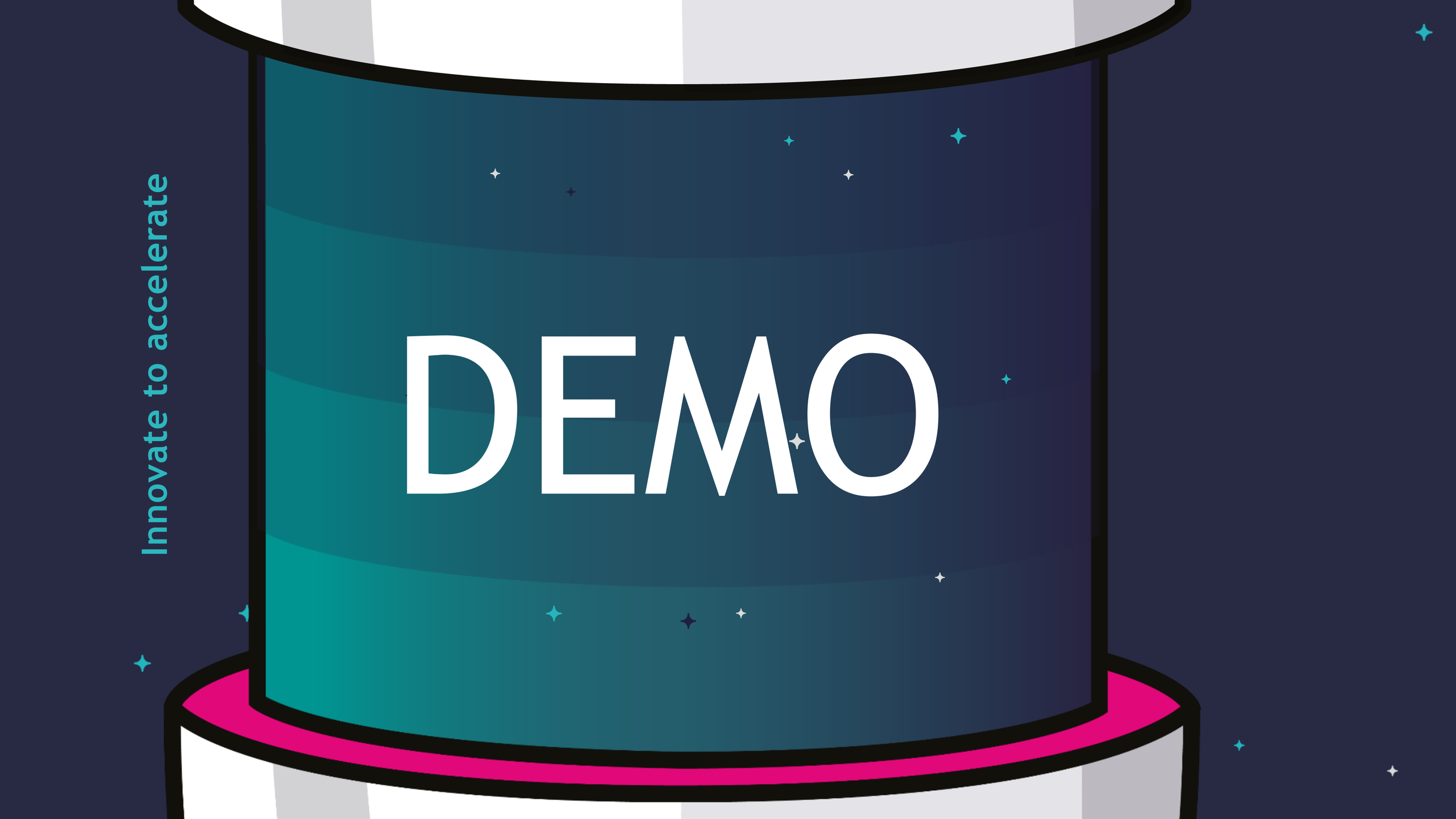
Out-of-the-Box Framework

- Execute pipeline
- Top
  - This pipeline will calculate the total number of objects (tables etc.) required to be copied in this run
- Middle
  - This pipeline will copy one batch of objects.
- Bottom
  - This pipeline will copy objects from one group



	Id	SourceObjectSettings	SourceConnectionSettingsName	CopySourceSettings	SinkObjectSettings	SinkConnectionSettingsName	CopySinkSettings
1	1	{ "schema": "Dimens...	NULL	{ "partitionOption": "Non...	{ "fileName": "DimensionCity.txt", ...	NULL	NULL
2	2	{ "schema": "Dimens...	NULL	{ "partitionOption": "Non...	{ "fileName": "DimensionCustomer.txt", ...	NULL	NULL
3	3	{ "schema": "Dimens...	NULL	{ "partitionOption": "Non...	{ "fileName": "DimensionDate.txt", ...	NULL	NULL
4	4	{ "schema": "Dimens...	NULL	{ "partitionOption": "Non...	{ "fileName": "DimensionEmployee.txt", ...	NULL	NULL
5	5	{ "schema": "Dimens...	NULL	{ "partitionOption": "Non...	{ "fileName": "DimensionPayment%20M...	NULL	NULL
6	6	{ "schema": "Dimens...	NULL	{ "partitionOption": "Non...	{ "fileName": "DimensionStock%20Item....	NULL	NULL
7	7	{ "schema": "Dimens...	NULL	{ "partitionOption": "Non...	{ "fileName": "DimensionSupplier.txt", ...	NULL	NULL
8	8	{ "schema": "Dimens...	NULL	{ "partitionOption": "Non...	{ "fileName": "DimensionTransaction%2...	NULL	NULL
9	9	{ "schema": "Fact", ...	NULL	{ "partitionOption": "Non...	{ "fileName": "FactMovement.txt", ...	NULL	NULL
10	10	{ "schema": "Fact", ...	NULL	{ "partitionOption": "Non...	{ "fileName": "FactOrder.txt", "fol...	NULL	NULL
11	11	{ "schema": "Fact", ...	NULL	{ "partitionOption": "Non...	{ "fileName": "FactPurchase.txt", ...	NULL	NULL
12	12	{ "schema": "Fact", ...	NULL	{ "partitionOption": "Non...	{ "fileName": "FactSale.txt", "fold...	NULL	NULL
13	13	{ "schema": "Fact", ...	NULL	{ "partitionOption": "Non...	{ "fileName": "FactStock%20Holding.txt...	NULL	NULL
14	14	{ "schema": "Fact", ...	NULL	{ "partitionOption": "Non...	{ "fileName": "FactTransaction.txt", ...	NULL	NULL

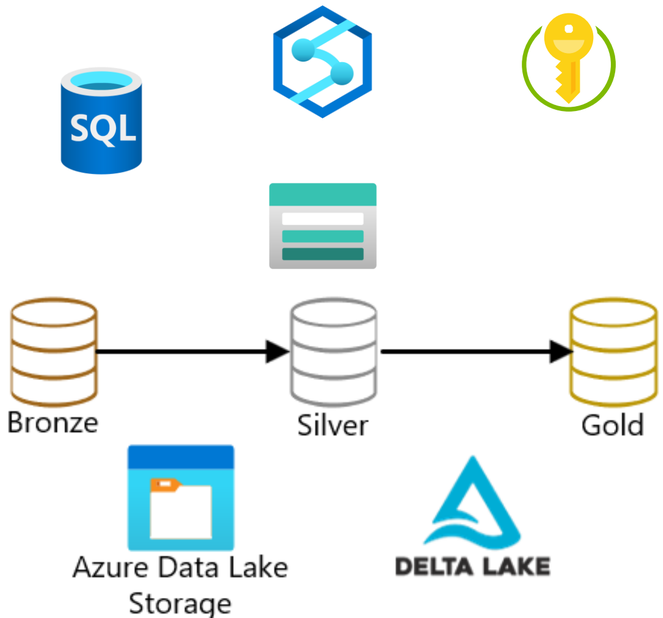
INSPARK


Innovate to accelerate
DEMO

Metadata-driven Framework

Custom-Made Framework

- Based on parameters
- Meta data => Azure SQL Database
- Azure Synapse Analytics and Azure Data Factory
- Azure Key Vault
- Azure Data lake
- Based on the Medaillon Architecture



Who is already using Parameters?

Metadata-driven Framework

Custom-Made Framework

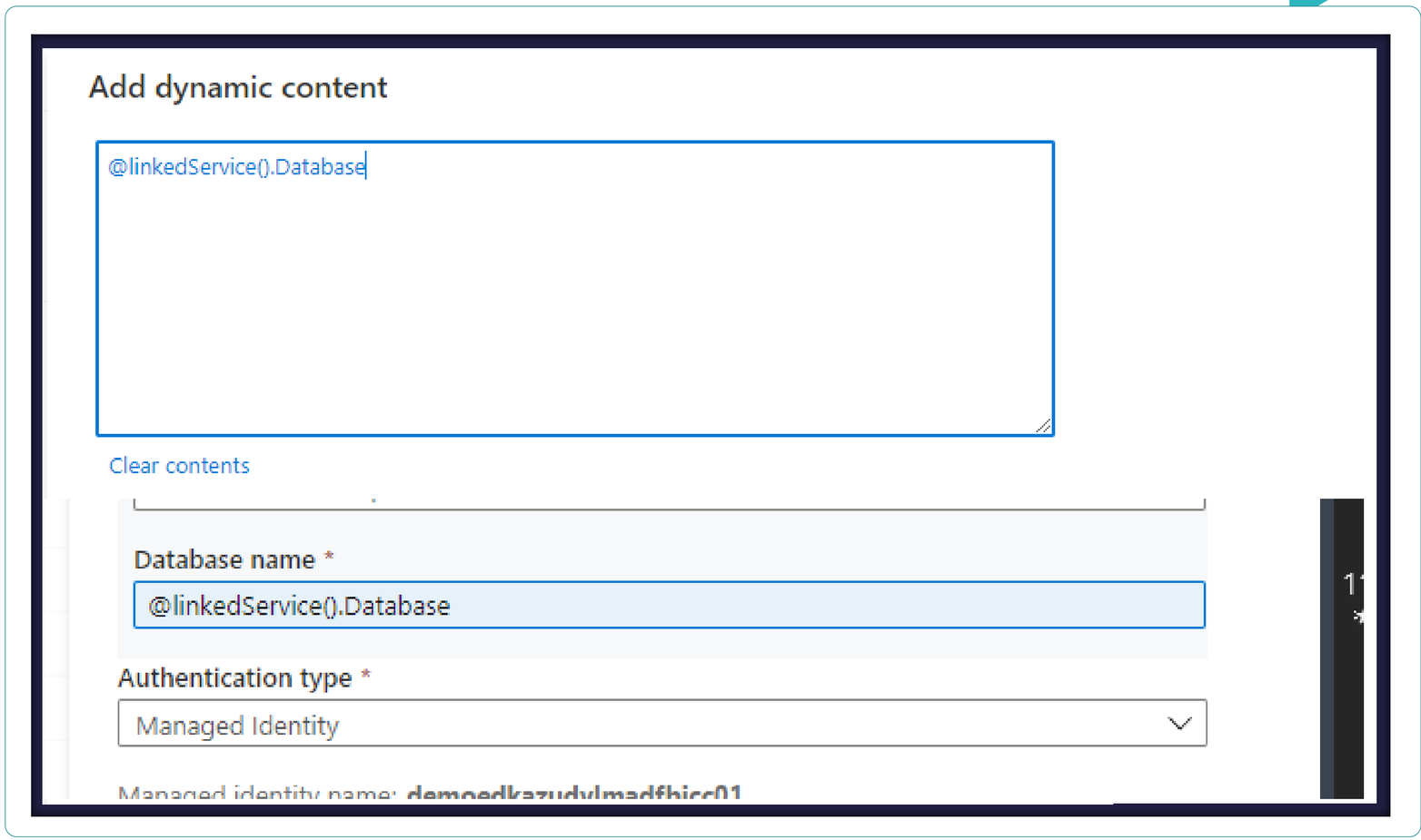
- Linked Services Parameter
- Dataset Parameter
- Dataflow Parameter
- Notebook Parameter
- Pipeline Parameter

INSPARK

Custom-Made Framework

Linked Services Parameters

- Connect to different Databases on same Server
- Connect to different Logical Servers based on the same Linked Service

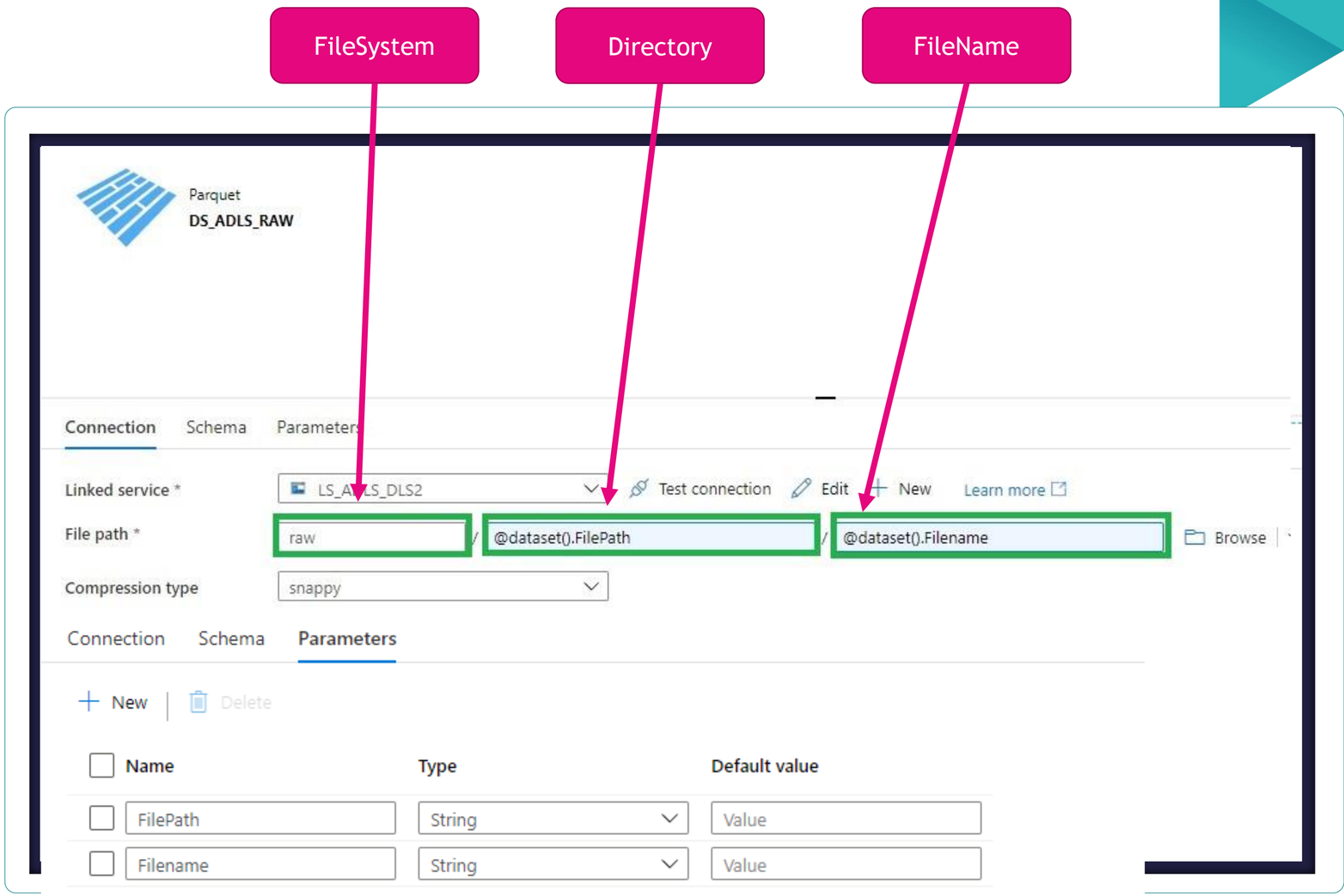


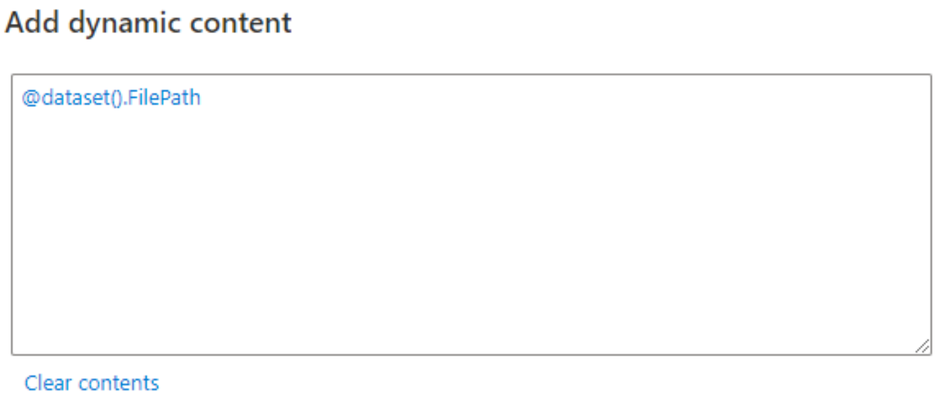
INSPARK

Custom-Made Framework

Dataset Parameters

- Create 1 dataset for all your Linked Services activities



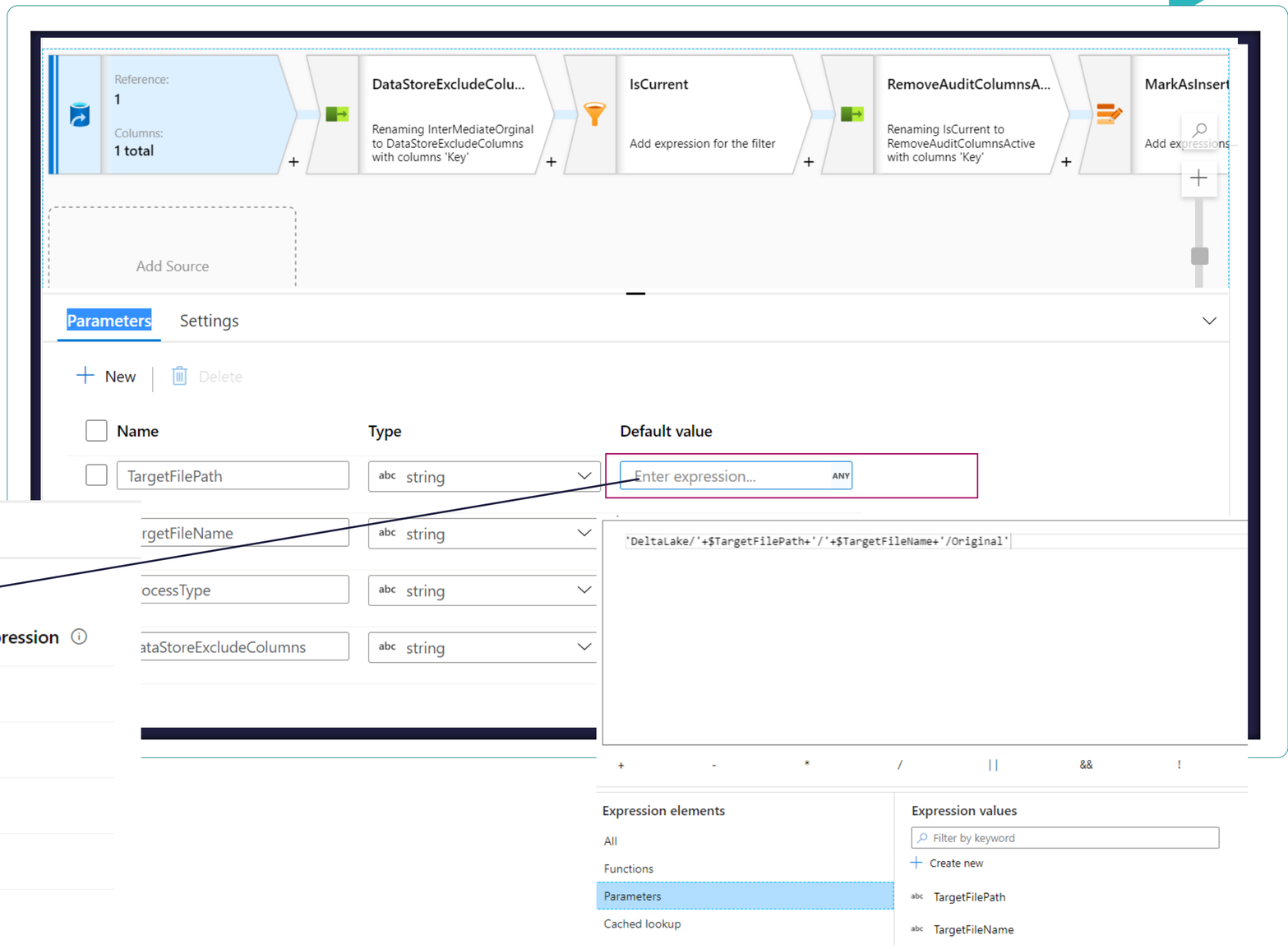


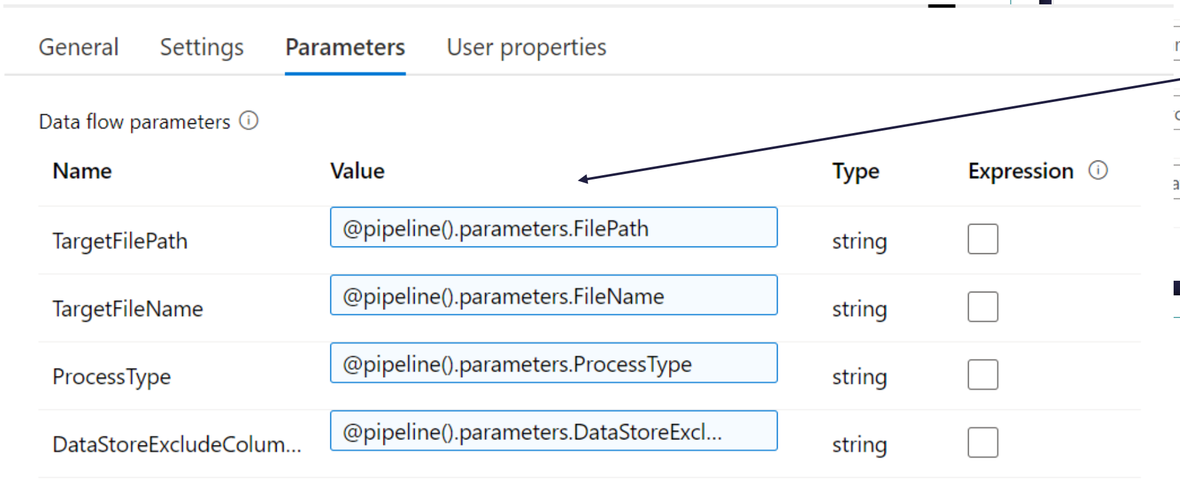
INSPARK

Custom-Made Framework

Dataflow Parameters

- Create a dynamically Dataflow
- Easy to re-use





Custom-Made Framework

Notebook Parameters

- Pass Parameters from Synapse/ADF Pipeline to Databricks
  - Using Widgets
- Pass Parameters from Synapse Pipeline to Synapse Notebook
  - Toggle parameter cell

```
# Creating widgets for leveraging parameters, and printing the parameters

datalake_connection ='labseuwdvlmdloxgn01.dfs.core.windows.net'
target_file =''
target_PKColumns ='ProductId'
source_file = 'AW/SalesLT_Product/2021/06/09/AW_SalesLT_Product_202106090916.parquet'
target_file_path =''
HashExcludeColumns =''
delta_db = 'delta'
IsIncremental =''
ReturnOutput = True
```

[2] ✓

Parameters

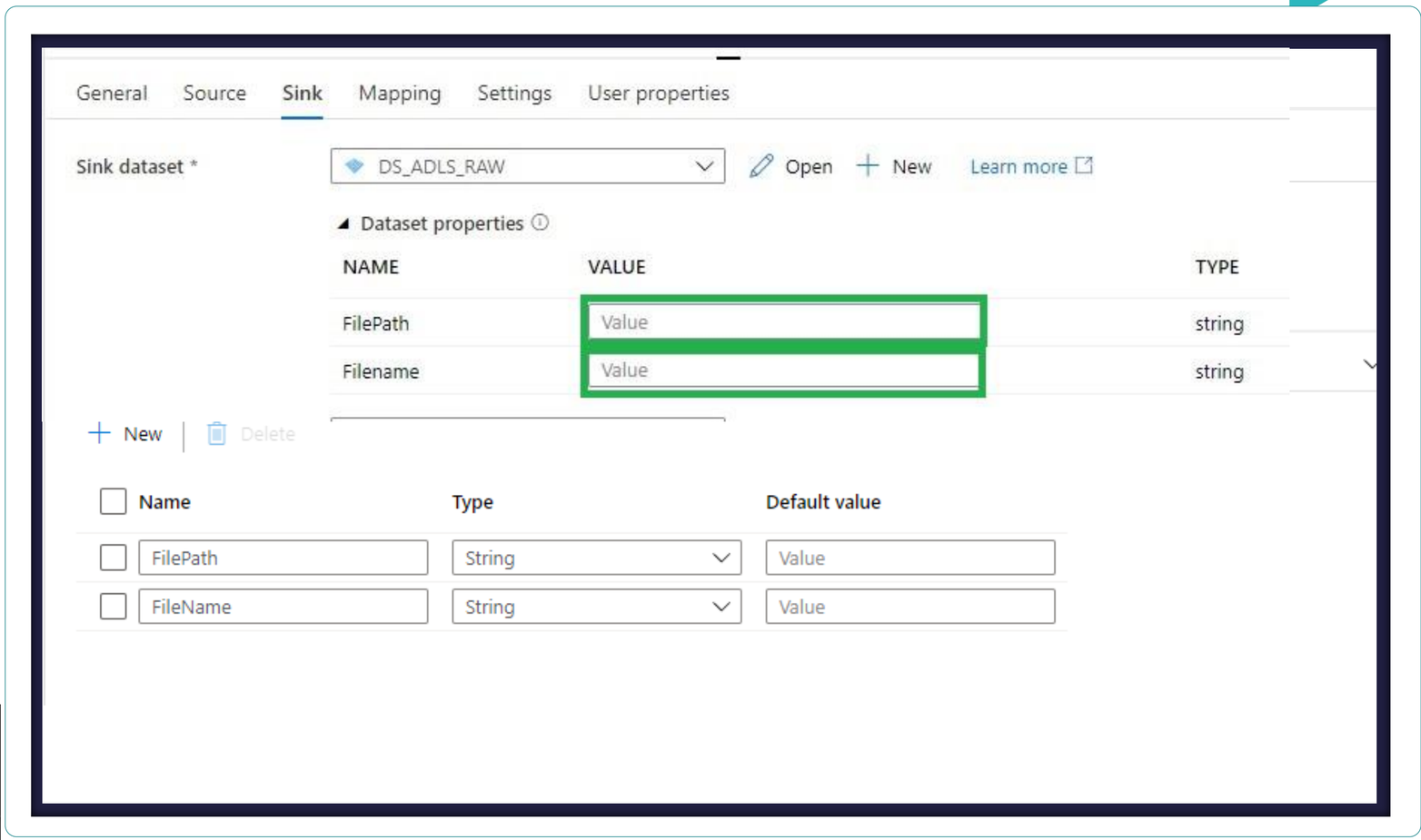
datalake_connection	@activity('Get EnvironmentSettings').o...
target_file	@{pipeline().parameters.TargetFileNa...
target_PKColumns	@{pipeline().parameters.PKColumns}
source_file	@{pipeline().parameters.FilePath}/@{pi...
target_file_path	@pipeline().parameters.TargetFilePath
HashExcludeColumns	@pipeline().parameters.HashExcludeC...

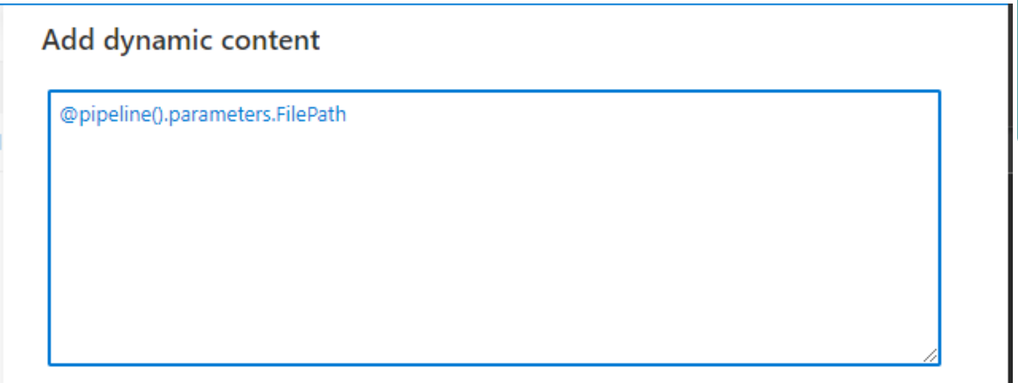
INSPARK

Custom-Made Framework

Pipeline Parameters

- Can be used across all your Pipelines
- Can easily control dataset, linked services and dataflows etc



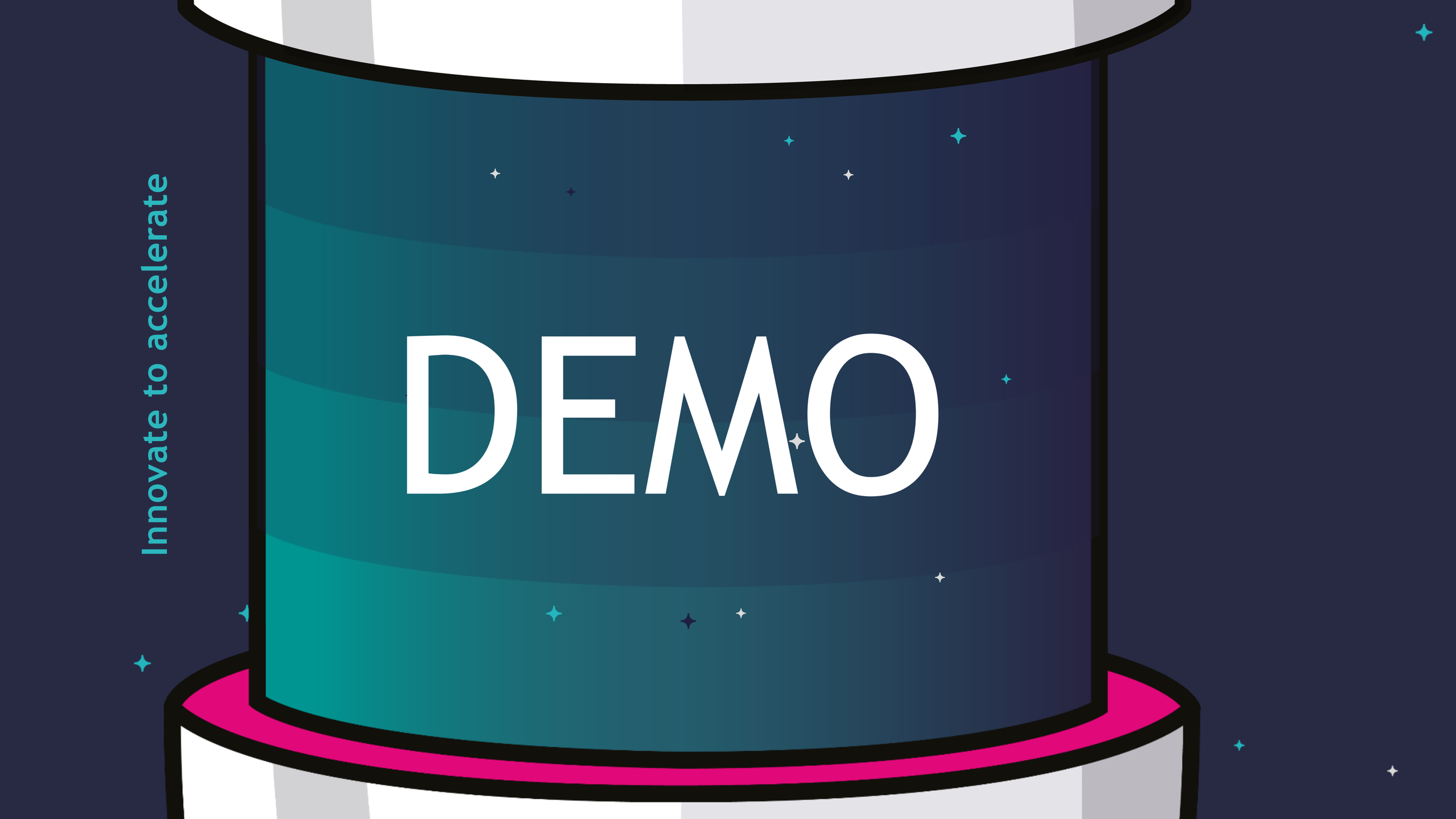


INSPARK

Parameters

- Linked Service
- DataSet
- Dataflow
- Notebooks
- Pipeline

INSPARK


Innovate to accelerate
DEMO

Custom-Made Framework

Can we build Synapse Pipelines dynamically?

Can we load the active(current) or historical records to a DataStore?

Can we extract data from my sources based on MetaData?

Can we build history from extracted data based on MetaData?

Can we log the execution of the Pipelines?

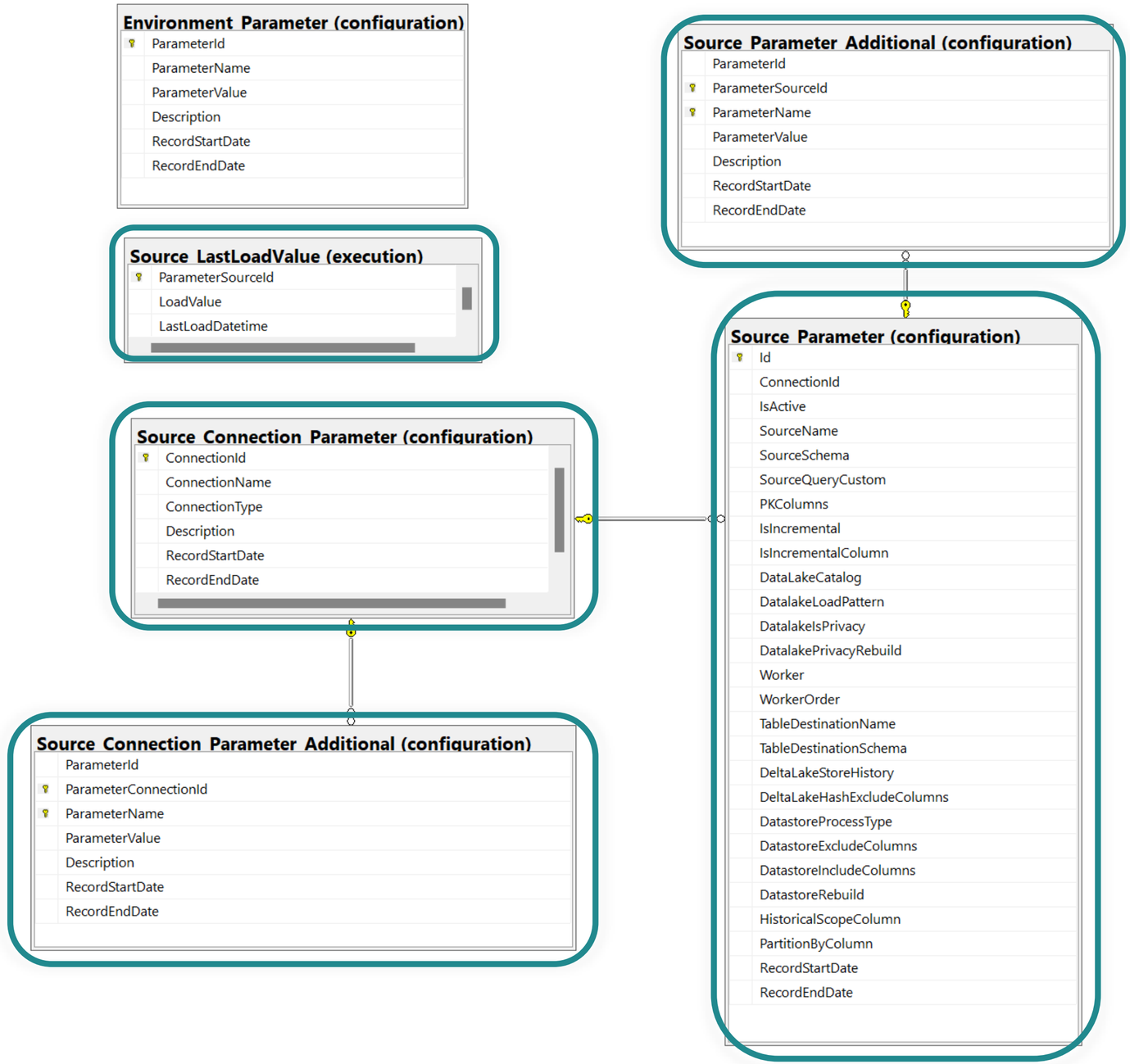
INSPARK


YES WE
CAN.

Custom-Made Framework

Configuration

- Source Parameters
  - Source Parameters Additional
- Source Connection Parameters
  - Source Connection Parameters additional
- Source Last LoadValue
- Versioning enabled (track changes)

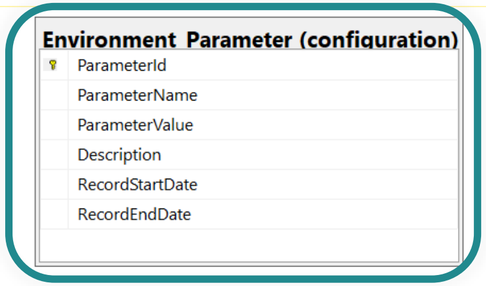


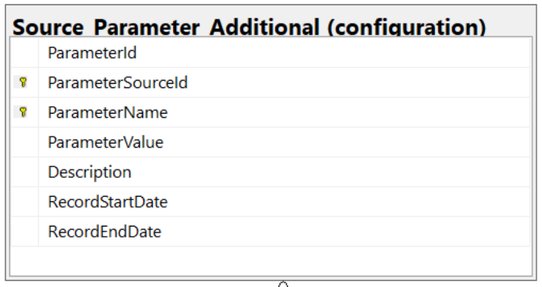
INSPARK

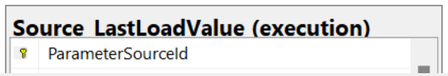
Custom-Made Framework

Environment

- Overarching settings that can be constantly reused.
- Global settings



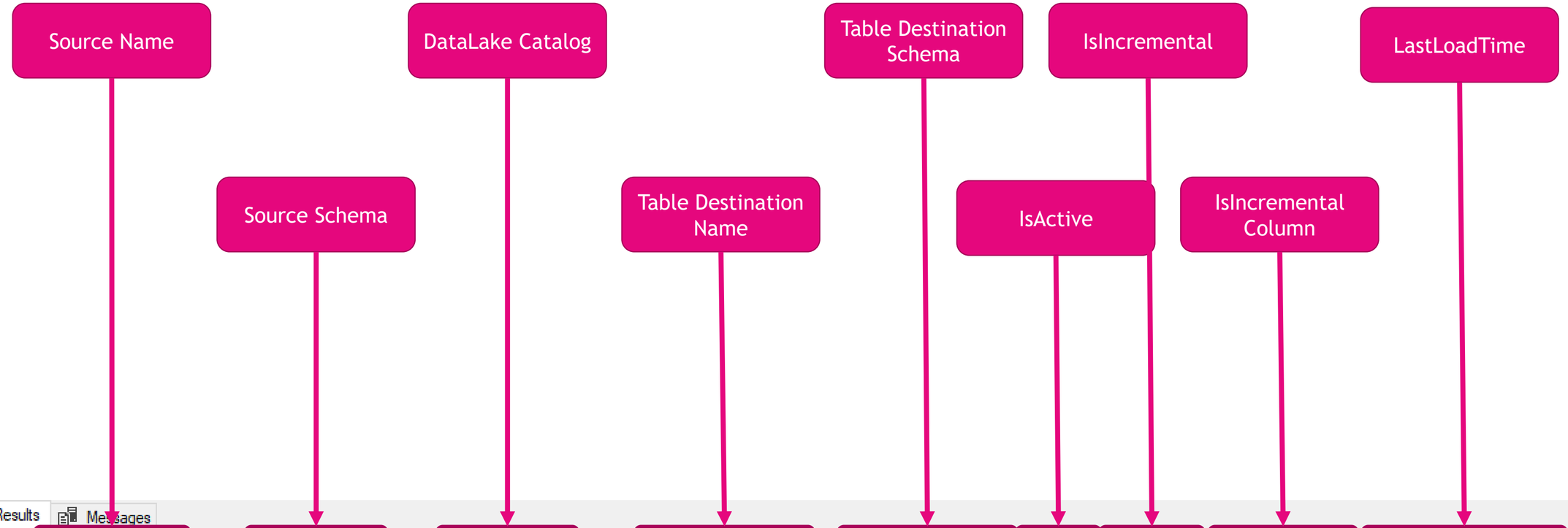




	ParameterId	ParameterName	ParameterValue	Description
1	1	Environment	DVLM	Enviroment short DVLM-TEST-ACPT-PROD
2	2	EnvironmentLong	development	Enviroment long develop-test-accpetance-production
3	3	AzureDataBricksClusterId	XXXX	ClusterId for Databricks Environment
4	4	ResourceGroupName	XXXX	ResourceGroupName of your Resources
5	5	KeyVaultConnection	https://edededeuwdvlmvaultoxgn01.vault.azure.net/	Connection to keyvault of your current environment
6	6	DataLakeConnection	https://edededeuwdvlmdloxgn01.dfs.core.windows.net/	DataLake Connection
7	7	AAS_ServerName	XXXX	Name of your Azure Analyis Service Server
8	8	SubscriptionId	xxxxx-xxxx-xxxx-xxxx-xxxxxxxx	SubscriptionId
9	9	TenantId	xxxxx-xxxx-xxxx-xxxx-xxxxxxxx	TenantId
10	10	KeyVaultConfigDB	ASQL-METADATAFRAMEWORK	Keyvault secret with Connection Details voor Config...
11	11	KeyVaultDatastore	ASQL-OXGN01-DATASTORE	Keyvault secret with Connection Details voor DataS...
12	12	IsDataStoreCSVEnabled	0	Do you want to create a CSV file from your DataSt...
13	13	IsDataStoreParquetEnabled	0	Do you want to create a Parquet file from your Data...
14	14	NumberOfDIUDataLake	4	DIU capacity for processing Source to DataLake av...
15	15	NumberOfDIUDataStore	0	DIU capacity for processing Deltalake to DataStore...
16	16	DataStoreSchema	Stg	Name of the schema which is used to Store your D...
17	17	IsSourceLoggingEnabled	1	Feature functionality to log ADF/Synapse
18	18	IsDataStoreLoggingEnabled	1	Feature functionality to log ADF/Synapse
19	19	IsDataStoreHistoricalEnabled	0	Are you building a Histroical Datastore model
20	20	IsDataVaultLoadEnabled	0	Are you building a DataVault model
21	21	IsHistoricalDatastoreLoadEnabled	0	Create Historical layer in SQL DataStore, loading D...
22	22	MaxRowsPerFile	1000000	Number of rows to split files
23	23	SynapseDedicatedSQLPool	DWH	Name of the Dedicated SQL Pool in Synapse
24	24	UseSynapseServerlessDataStore	1	Use Serverless Engine to read from DeltaLake to D...
25	25	DeltaLakeModule	2	0=DataBricks 1=DataFlow 2=SynapseSpark
26	26	DatastoreModule	1	1=Use datastore 0=No datastore
27	27	LakehouseModule	1	1=Use lakehouse 0=No lakehouse
28	28	TeamsWebhook		Webhook to sent general pipeline messages to

INSPARK

Custom-Made Framework



Results | Messages

	Id	SourceName	SourceSchema	DataLakeCatalog	TableDestinationName	TableDestinationSchema	IsActive	IsIncremental	IsIncrementalColumn	LastLoadtime
1	1	PurchaseOrderLines	Purchasing	WWI	PurchaseOrderLines	Purchasing	1	1	LasteditedWhen	2021-01-07 13:00:00.000
2	2	PurchaseOrders	Purchasing	WWI	PurchaseOrders	Purchasing	1	1	LasteditedWhen	2021-01-07 13:00:00.000
3	3	SupplierTransactions	Purchasing	WWI	SupplierTransactions	Purchasing	1	1	LasteditedWhen	2021-01-07 07:00:00.000
4	4	CustomerTransactions	Sales	WWI	CustomerTransactions	Sales	1	1	LasteditedWhen	2021-01-07 12:00:00.000

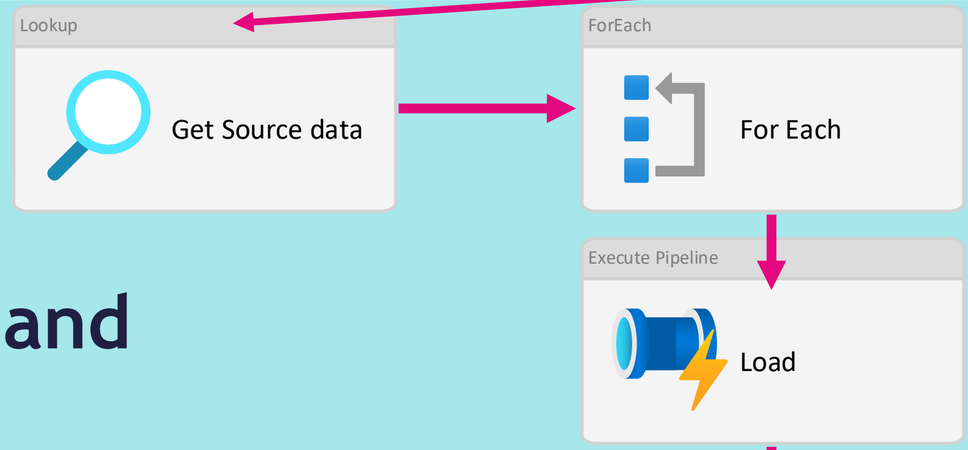
INSPARK

Custom-Made Framework

Pipelines

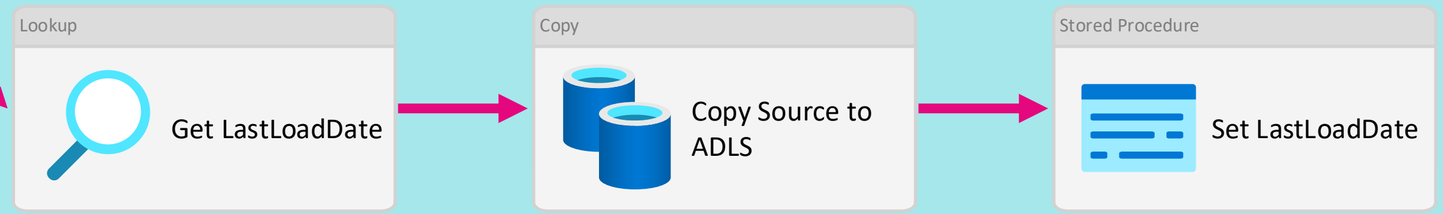
```
SELECT [PipelineParameterId]
      ,[SourceName]
      ,[SourceSchema]
      ,[SelectQuery]
      ,[SelectLastLoaddate]
      ,[FilePath]
      ,[FileName]
      ,[TableDestinationName]
      ,[ProcessType]
      ,[IsActive]
      ,[IsIncremental]
      ,[IsIncrementalColumn]
      ,[LastLoadtime]
  FROM [execution].[Pipeline_DataLake_Files]
```

Command



Execute

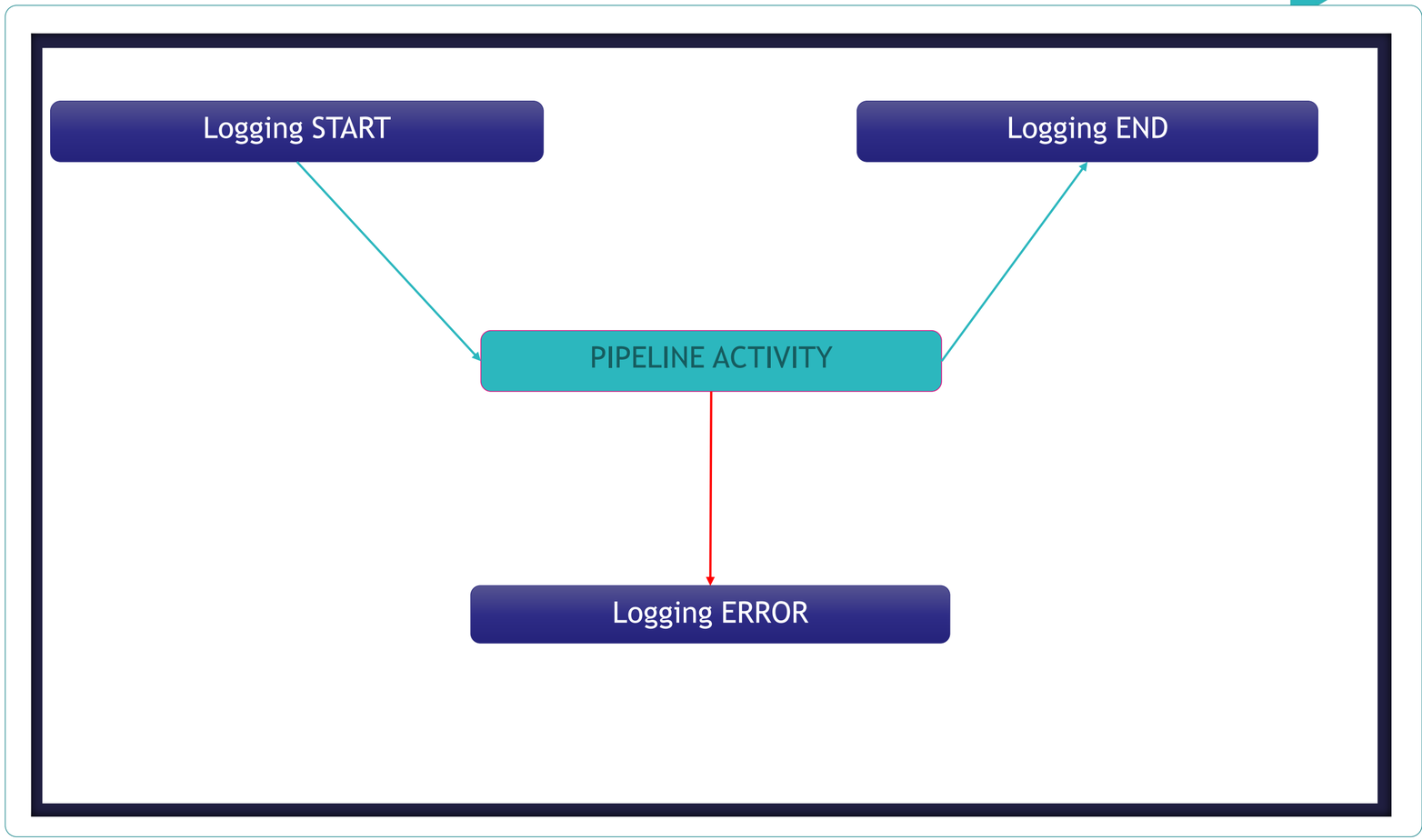
```
SELECT case when 1=1 then
convert(varchar,max(LasteditedWhen),120) else
convert(varchar,getdate(),120)  end as LastLoadDate
FROM SourceSchema.SourceTable
```



Custom-Made Framework

Logging

- Log Start and End Time of records
- Log Extracted Records
- Log Execution Failure
- Create Pipeline_ExecutionLog table



INSPARK

Custom-Made Framework

Logging

- Log Start and End Time of records
- Log Extracted Records
- Log Execution Failure
- Create Pipeline_ExecutionLog table

BEGIN
Insert new Record
Insert Metadata
Insert Start time

END
End Time
Status(1)
Row Counts
Pipeline Details

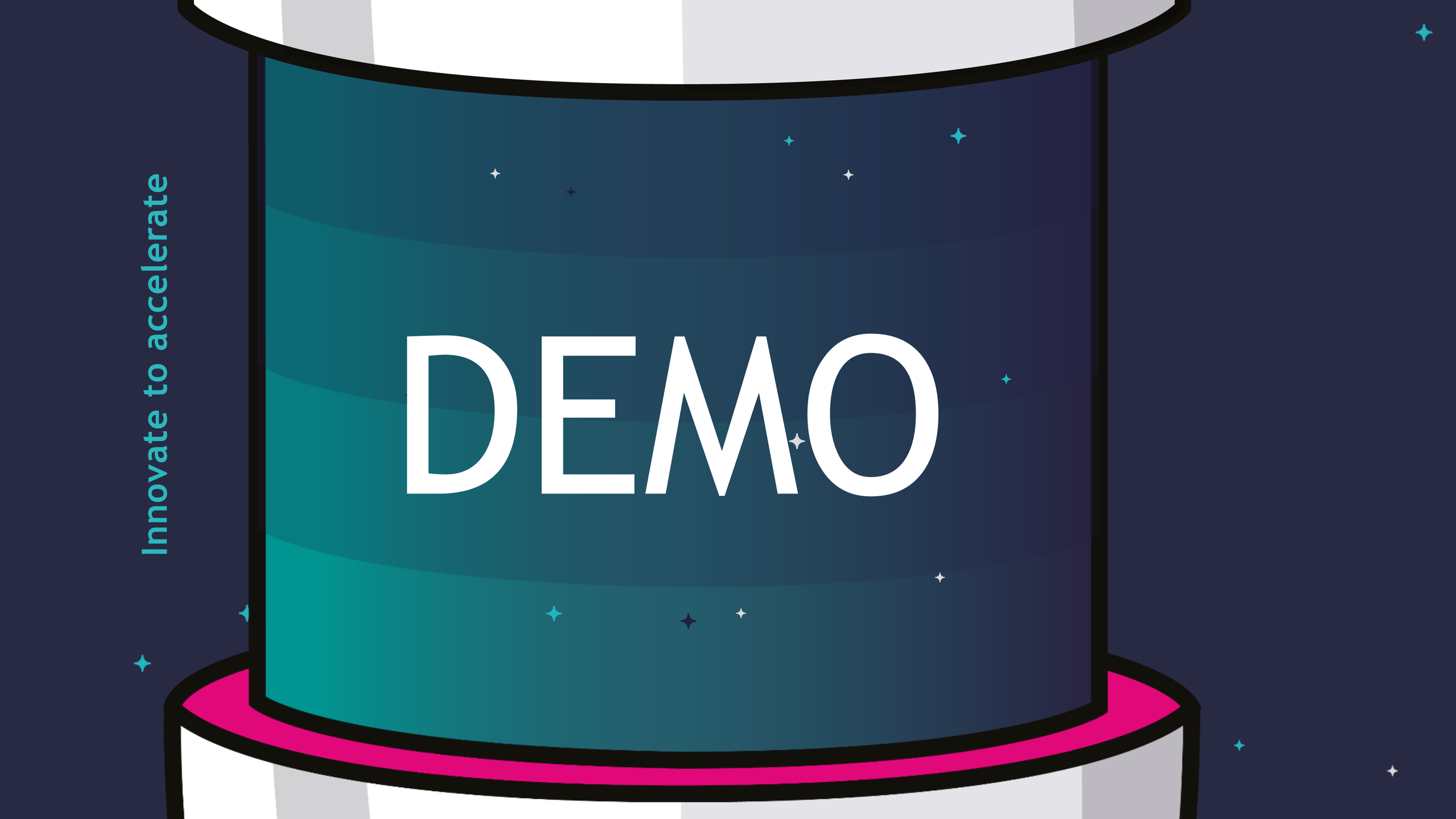
ERROR
End Time
Status(2)
Failure Message

Pipeline_ExecutionLog

Results | Messages

	LogId	StartTime	EndTime	Status	RowsRead	RowsNew	RowsUnchanged	RowsDeleted	RowsSkipped	Reference	NumberOfParallelCopies	IntegrationRuntime	DataConsistencyVerification	ReportLineageToCatalog
1	5	2021-02-08 19:55:35.517	2021-02-08 19:55:57.687	1	84	84	0	0	0	DataMovement	1	DefaultIntegrationRuntime (West Europe)	Verified	N/A
2	4	2021-02-08 19:55:04.670	2021-02-08 19:55:28.327	1	2	2	0	0	0	DataMovement	1	DefaultIntegrationRuntime (West Europe)	Verified	N/A
3	3	2021-02-08 19:54:35.920	2021-02-08 19:54:58.937	1	2	2	0	0	0	DataMovement	1	DefaultIntegrationRuntime (West Europe)	Verified	N/A
4	2	2021-02-08 19:54:09.117	2021-02-08 19:54:29.653	1	9	9	0	0	0	DataMovement	1	DefaultIntegrationRuntime (West Europe)	Verified	N/A
5	1	2021-02-08 19:53:57.763	NULL	0	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

	LogId	StartTime	En...	Status	FailureMessage
1	15	2021-02-0...	20...	1	NULL
2	14	2021-02-0...	20...	1	NULL
3	13	2021-02-0...	20...	2	Failure happened on 'Source' side. ErrorCode=SqlOperationFailed,'Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException,Message=A database operation failed with the followin...
4	12	2021-02-0...	20...	1	NULL
5	11	2021-02-0...	20...	2	NULL


Innovate to accelerate
DEMO

Custom-Made Framework

Can we build Synapse Pipelines dynamically?

Can we load the active(current) or historical records to a Lakehouse?

Can we extract data from my sources based on MetaData?

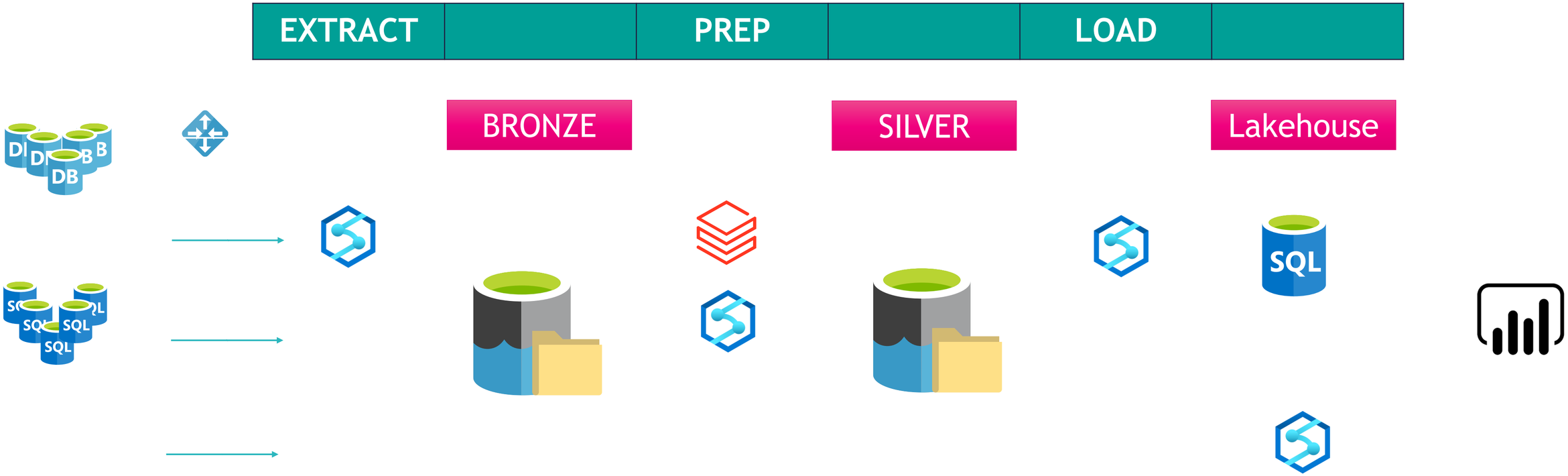
Can we build history from extracted data based on MetaData?

Can we log the execution of the Pipelines?

INSPARK

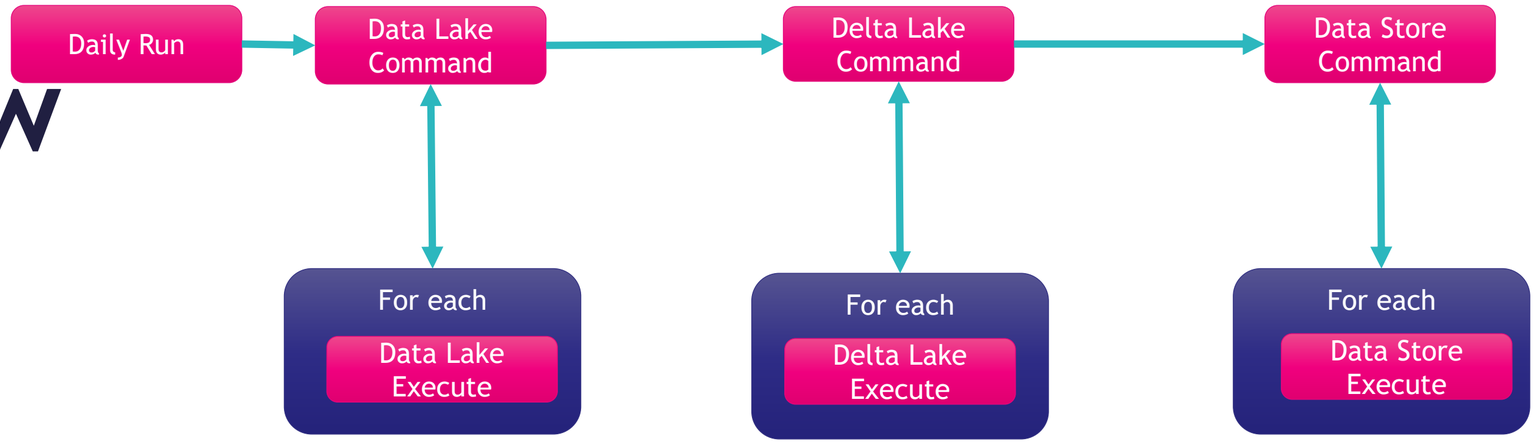
Custom-Made Framework

Loading Pattern

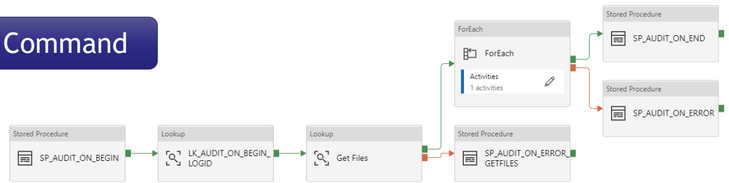


Custom-Made Framework

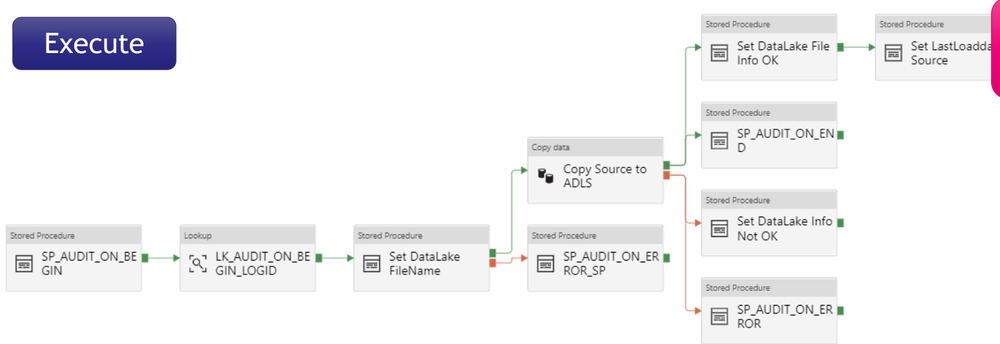
Process Flow

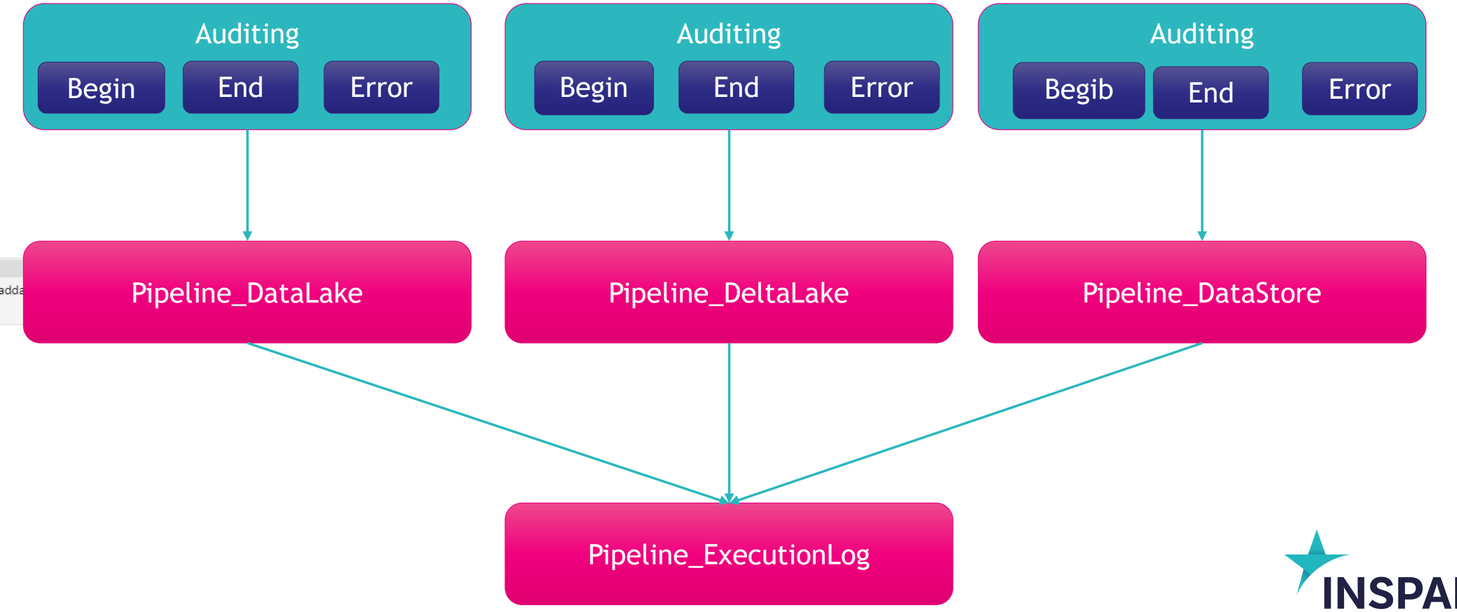


Command

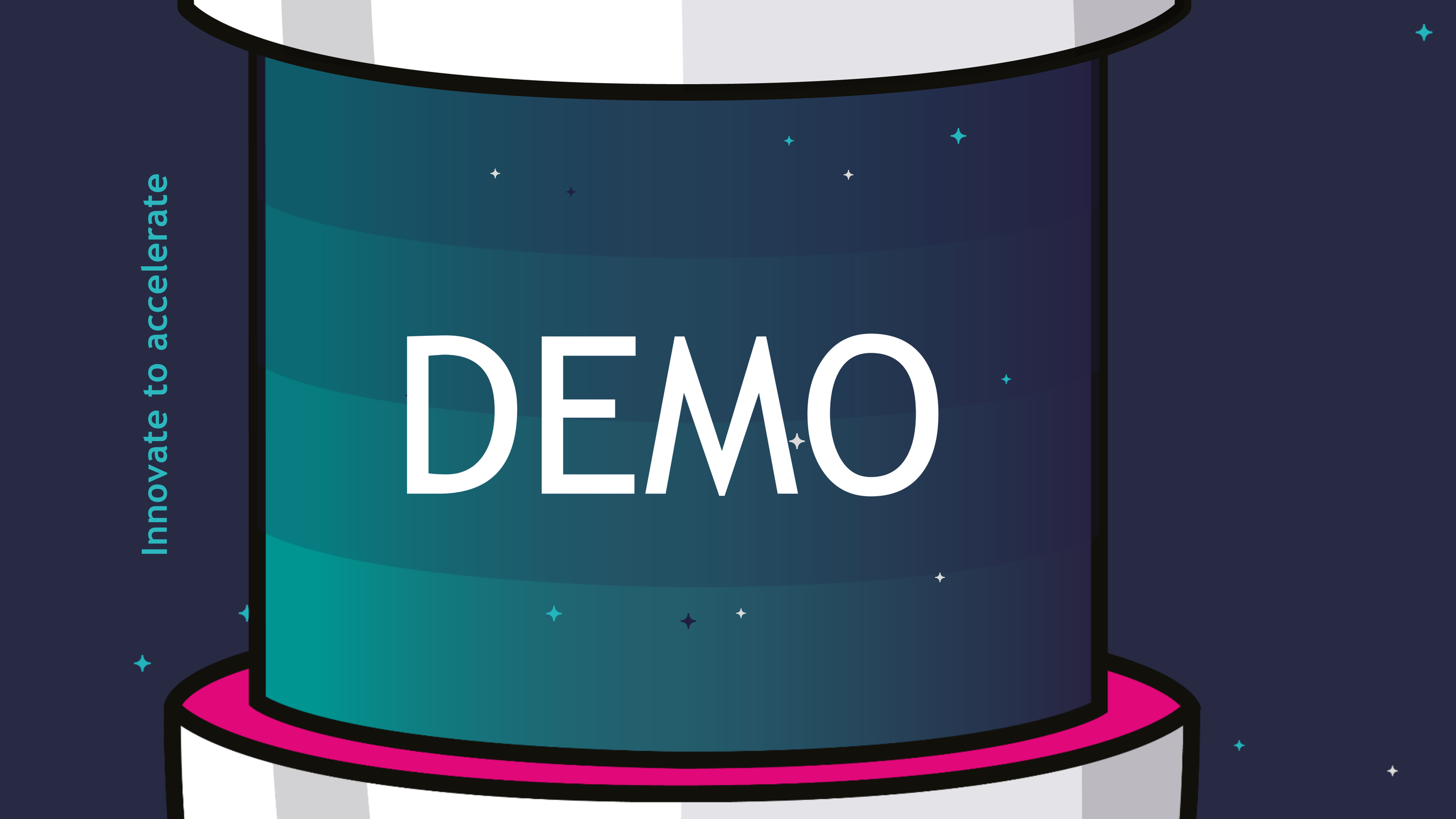


Execute





INSPARK


Innovate to accelerate
DEMO

Recap

- Can we build Synapse Pipelines dynamically? ✓
- Can we load the active(current) or historical records to a Lakehouse? ✓
- Can we extract data from my sources based on MetaData? ✓
- Can we build history from extracted data based on MetaData? ✓
- Can we log the execution of the Pipelines? ✓

Questions?

INSPARK

Erwin de Kreuk

Principal Consultant | Lead Data and AI | Data Platform MVP | Public Speaker …

Erwin de Kreuk

Principal Consultant – Lead Data & AI
InSpark

@erwindekreuk

linkedin.com/in/erwindekreuk

erwindekreuk.com github.com/edkreuk

Microsoft® Most Valuable Professional