INSPARK

Drive Efficiency and Reliability Using Metadata-Driven Frameworks

09-10-2024

Erwin de Kreuk

Marco Hansma

We Are InSpark

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

INSPARK

Thank you, partners

element 61
experience & expertise

Cloubis

Microsoft

MICROPOLE BELUX

PLAINSIGHT

Adapt and Enable

Tabular Editor

delaware

datashift

U2U

cegeka

EpicData.

LACO

tillit data shapers

ilytix

bmatix
Act informed

randstad digital

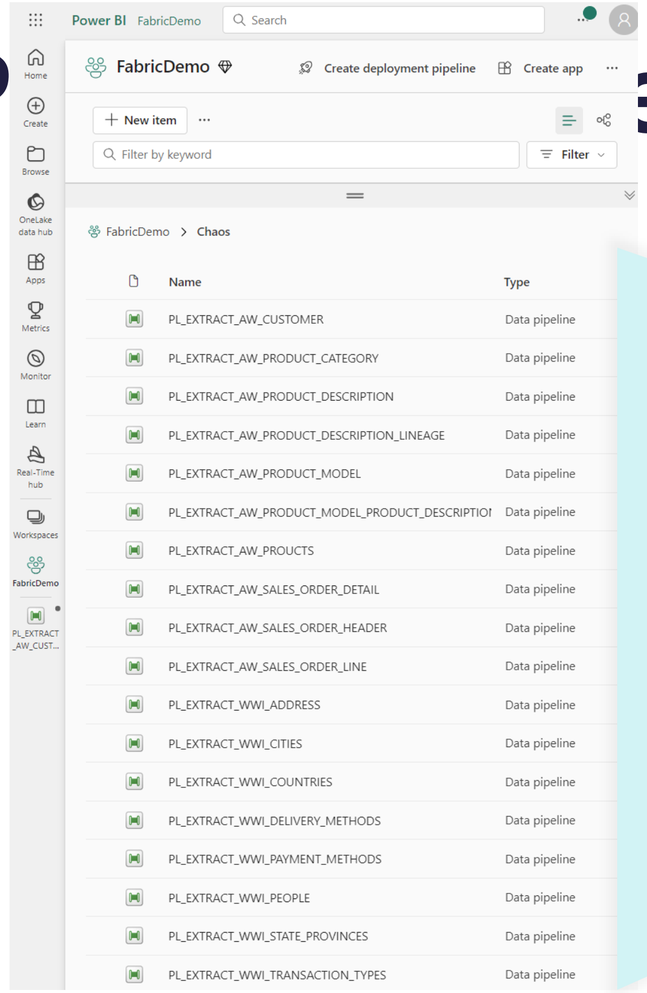
KOHERA

MONIN
Database Managed Services

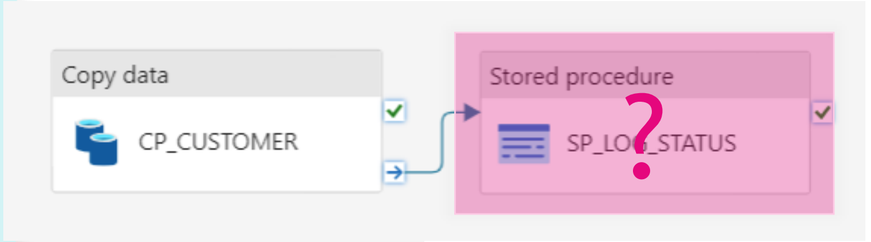
inetum realdolmen
Positive digital flow

TriFinance
BEYOND ADVISORY

P ement



Content of each pipeline



INSPARK

Erwin de Kreuk

Principal Consultant

Lead Data & AI InSpark

Let’s connect

Microsoft® Most Valuable Professional

@erwindekreuk

linkedin.com/in/erwindekreuk

erwindekreuk.com

github.com/edkreuk

https://sessionize.com/erwin-de-kreuk/

Marco Hansma

Architect InSpark

Let’s connect

linkedin.com/in/marcohansma

Objectives

Data platform challenges

Why

Medallion Architecture

Data Pipeline Parameters

Meta Data DrivenFramework

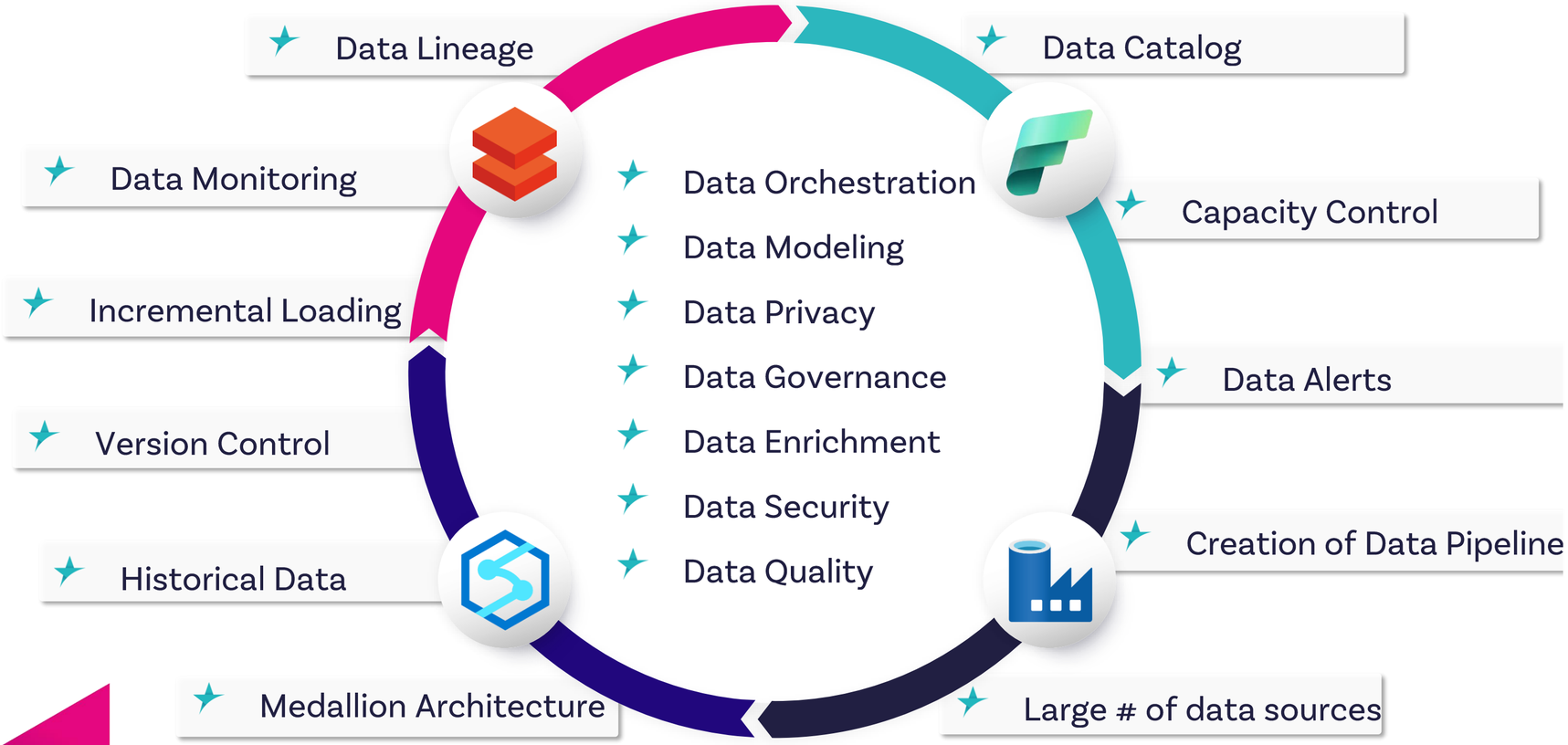
Data Pipeline Logging

Recap

INSPARK

Data platform Challenges

'From data source to data model' to report



“Simplify,

I am the Chief Data Information officer and don’t want to be the Chief Integration Officer

Help me to simplify and automate my Data Orchestration.”

Every CDO, Every organisation

What would help

- Simplicity in connecting data sources
- First results within hours
- Focus on business value instead of data integration
- Meta Data Driven
  - Standardized data pipelines, Notebooks, orchestration and Way of Work
- Overview of data process flows
- Detailed logging information
- Integration of solution like:
  - Data Privacy
  - Data Quality
  - Data Governance
- From data “Spaghetti to Lasagna”
  - Building a uniform data architecture

I am the Chief Data Information Officer and don’t want to be the Chief Integration Officer

Help me to simplify and automate my Data Orchestration

Why

Scalability

Automation

Traceability

Flexibility

INSPARK

Out-of-the-Box Framework

Ready-to-use.

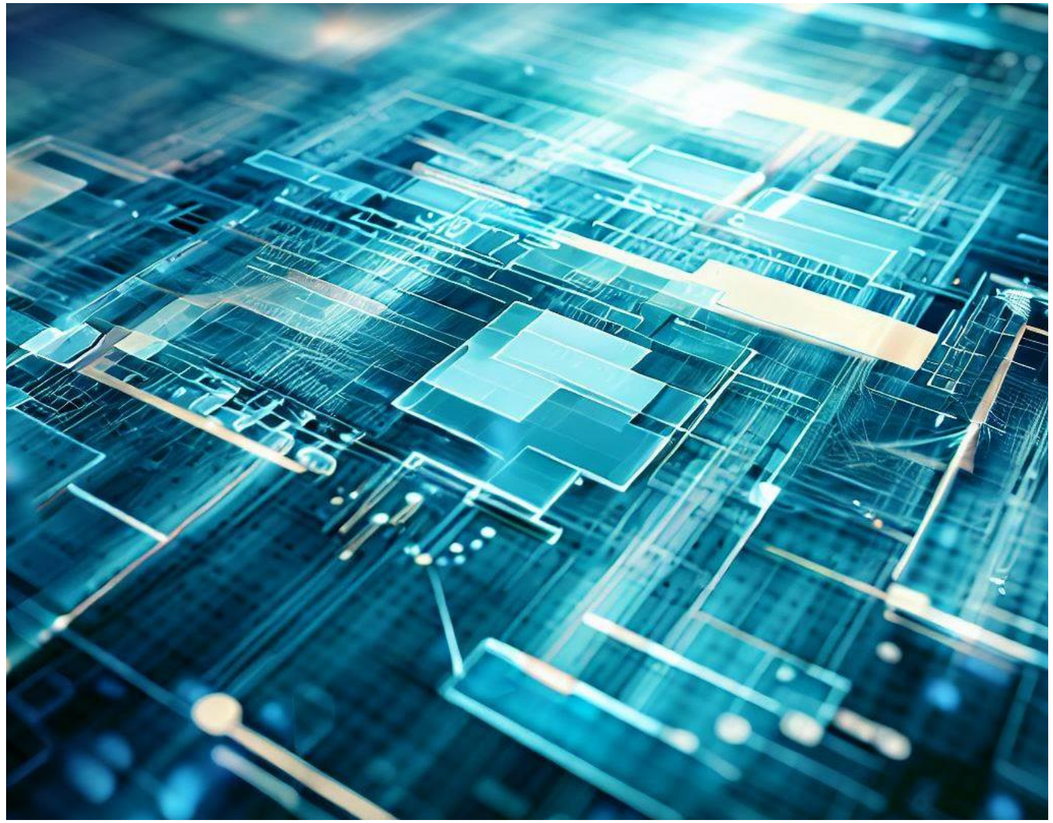
Rapid implementation.

Limited customization.

Lower development effort.

Lower upfront costs.

Ongoing support and updates.



INSPARK

Custom-Made Framework

Tailored to specific needs.

Full control over design and features.

Higher development effort.

Flexibility and extensibility.

Higher upfront costs.

INSPARK

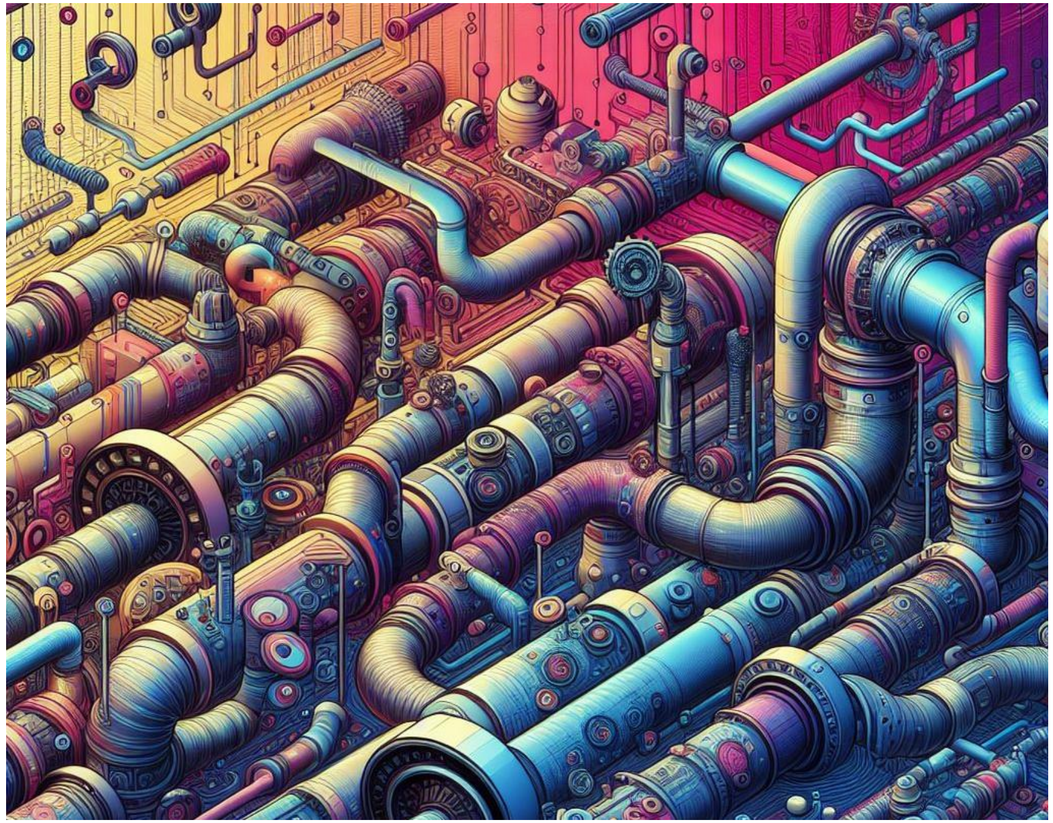
Custom-Made Framework

Based on parameters

Meta data => Azure SQL Database / Json / ……

Microsoft Fabric but also on Azure Synapse Analytics and Azure Data Factory

Based on the Medallion Architecture

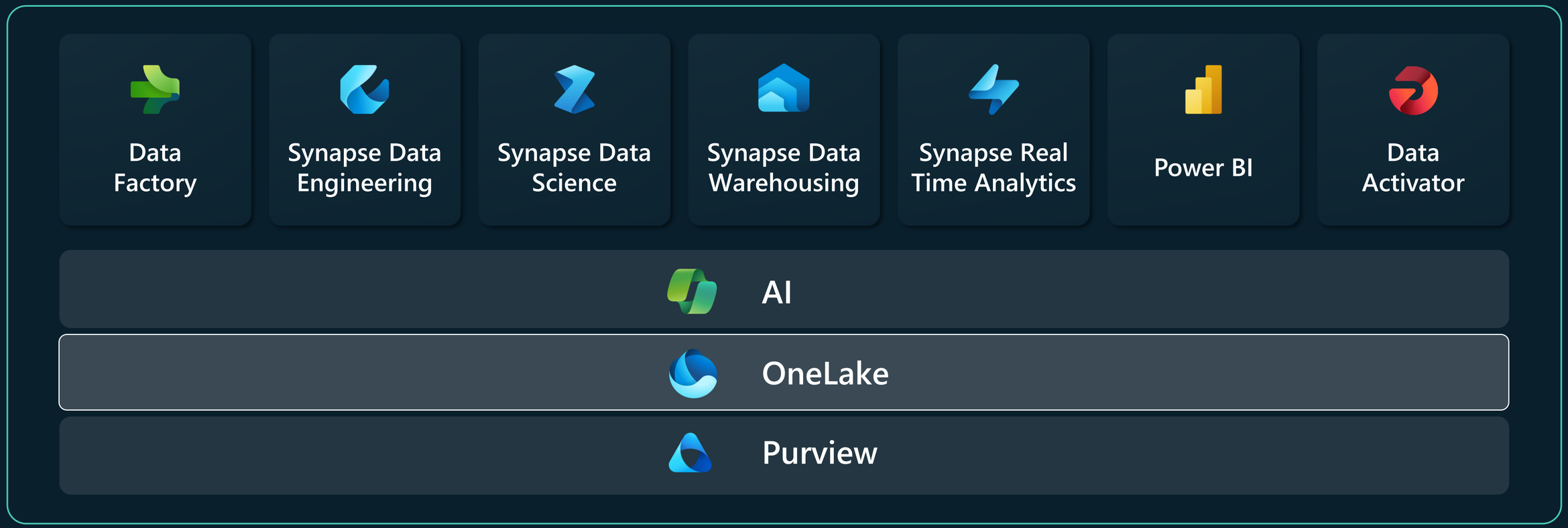


INSPARK

Who is already using Parameters?

Microsoft Fabric

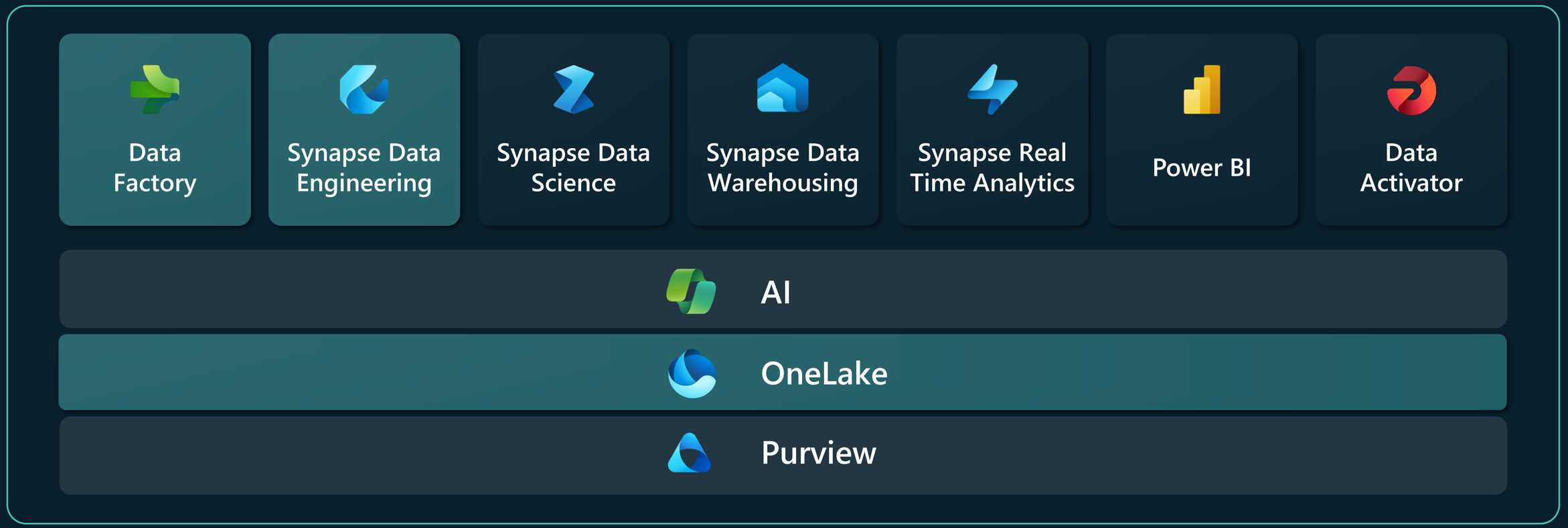
The unified data platform for the era of AI



Unified architecture | Unified experience | Unified governance | Unified business model

Microsoft Fabric

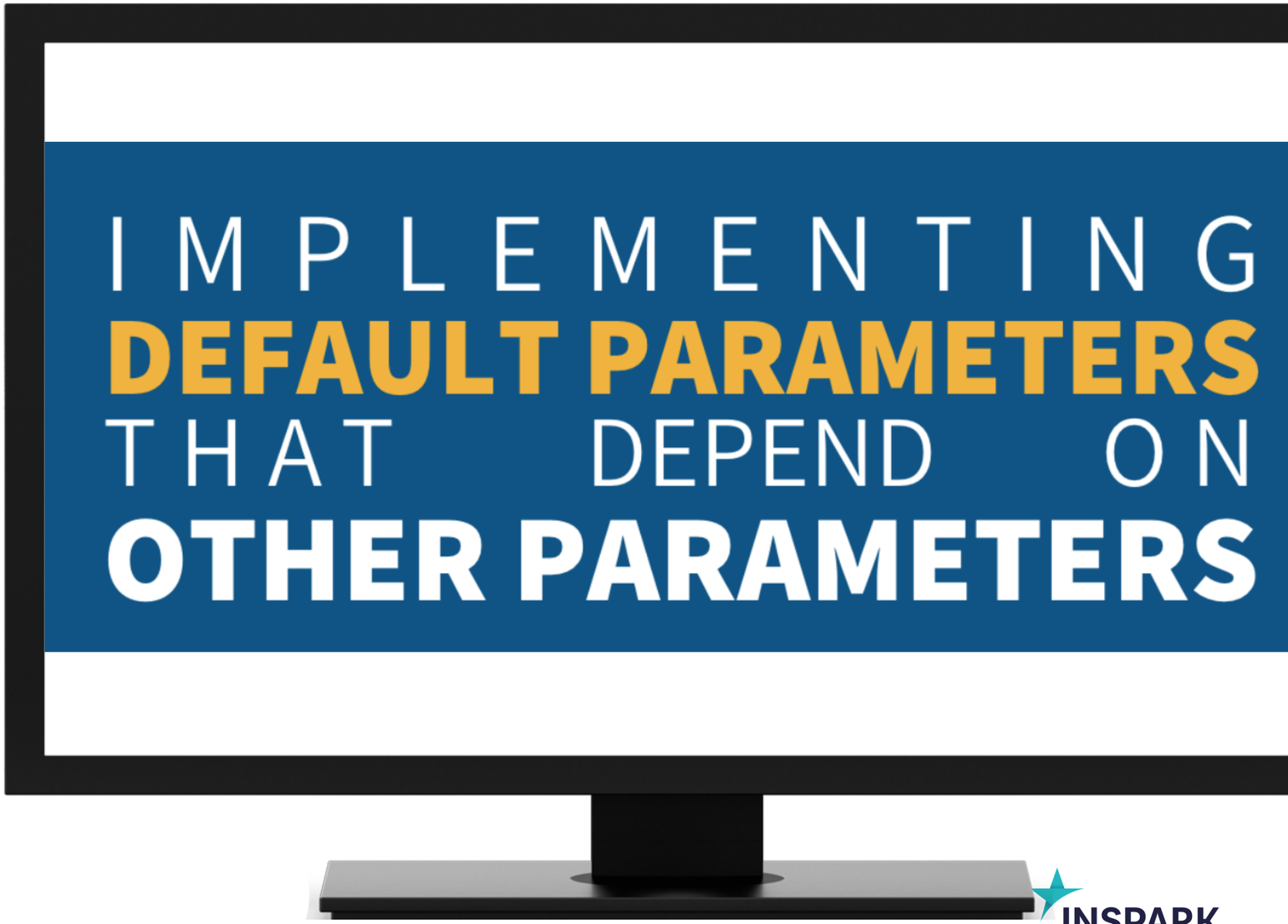
The unified data platform for the era of AI



Unified architecture | Unified experience | Unified governance | Unified business model

Parameters

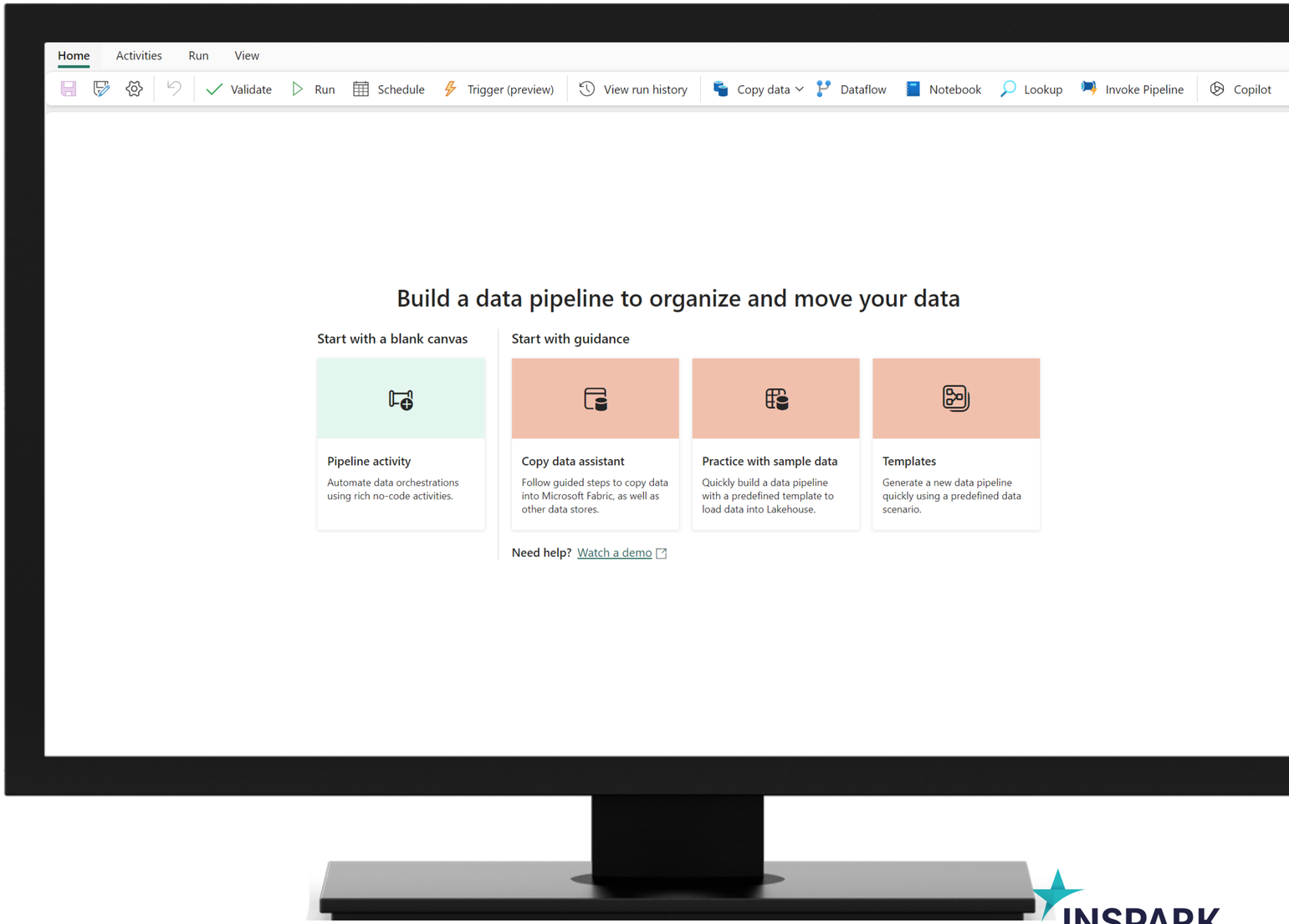
- Pipeline Parameters
  - Copy Activity Parameters
  - Foreach Parameters
- Notebook Parameters



INSPARK

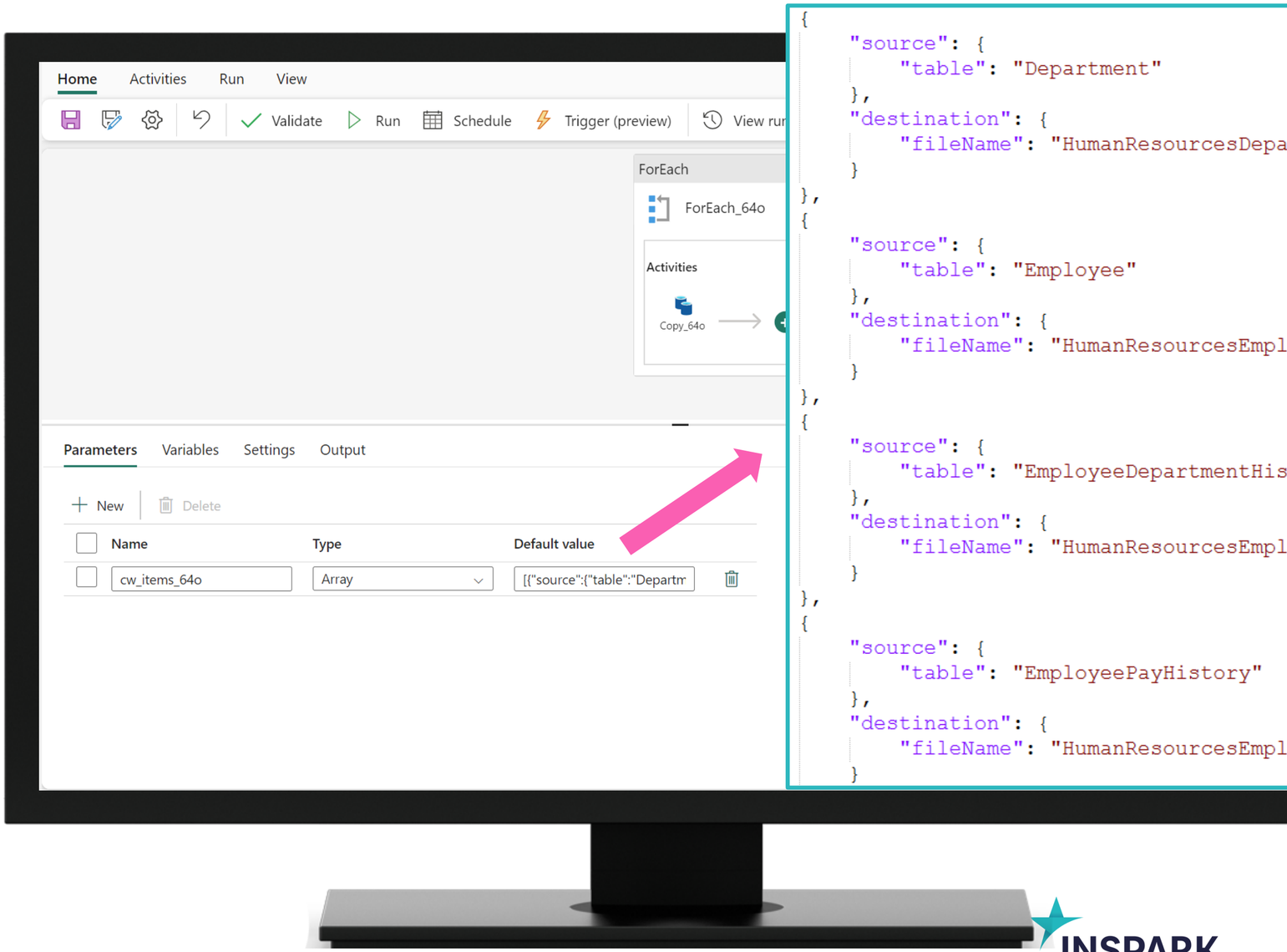
Copy Assistant

- Reads data from a source data store.
- Performs serialization/deserialization, compression/decompression, column mapping, and so on.
- Writes data to the destination data store.



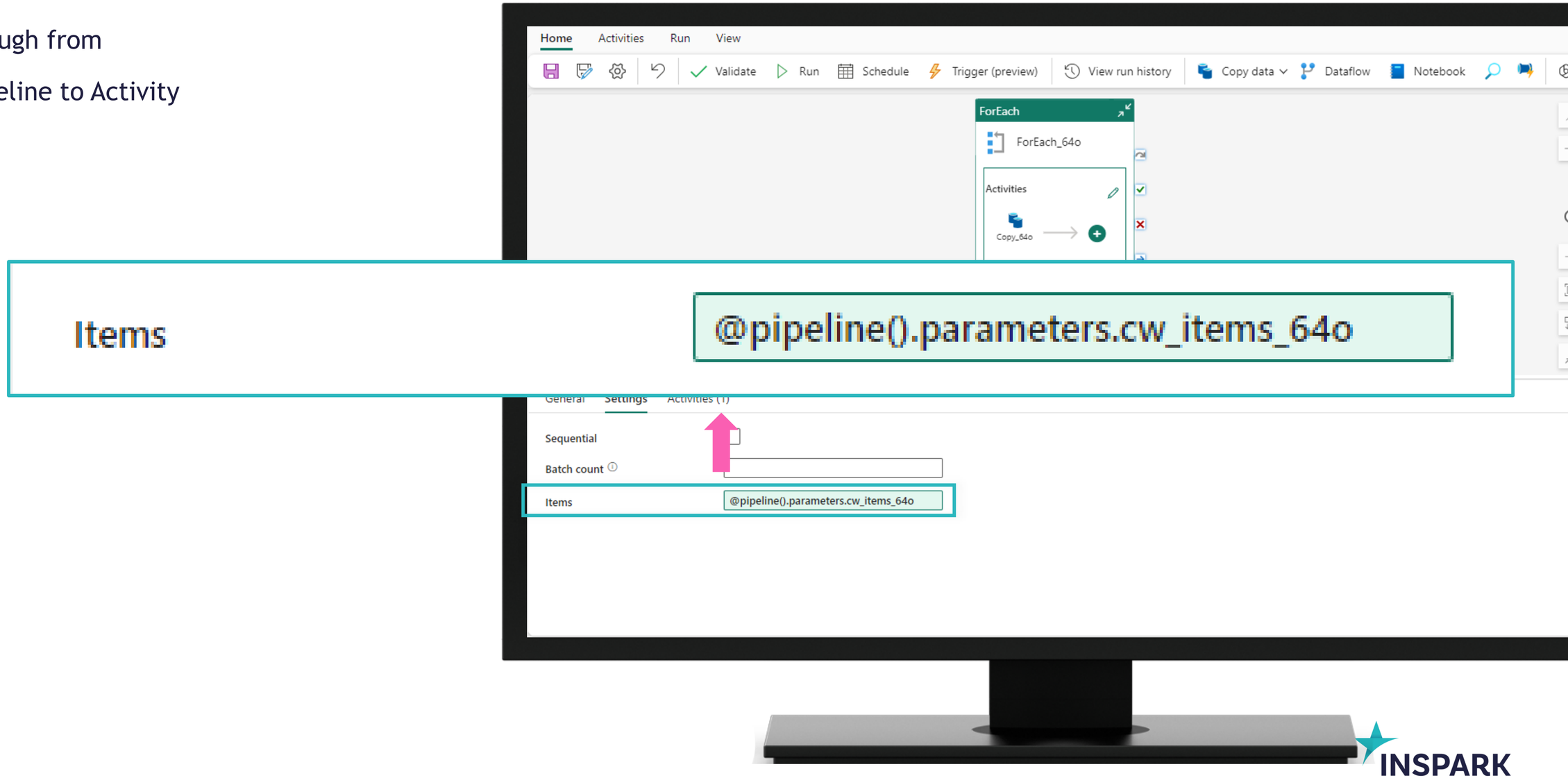
Understand Parameters

- Copy data assistant
  - Select Data Source
  - Select Tables
  - Select Destination
  - Select Filetypes
  - Review actions



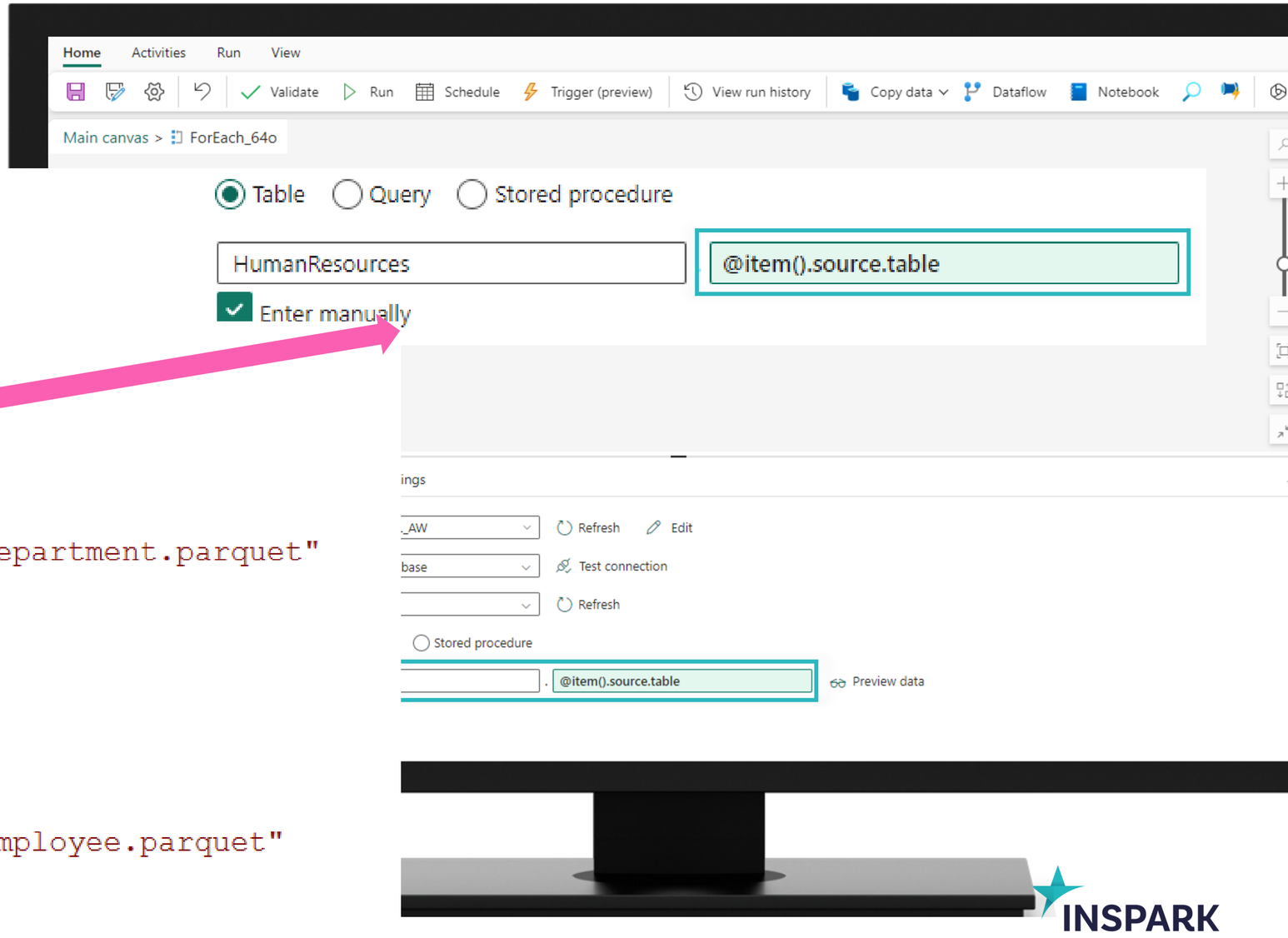
Pipeline Parameters

- Pass through from
  - Pipeline to Activity



Copy Activity Parameters

- Pass Parameters from Pipeline to Copy activity
- Use Parameters from For Each Activity

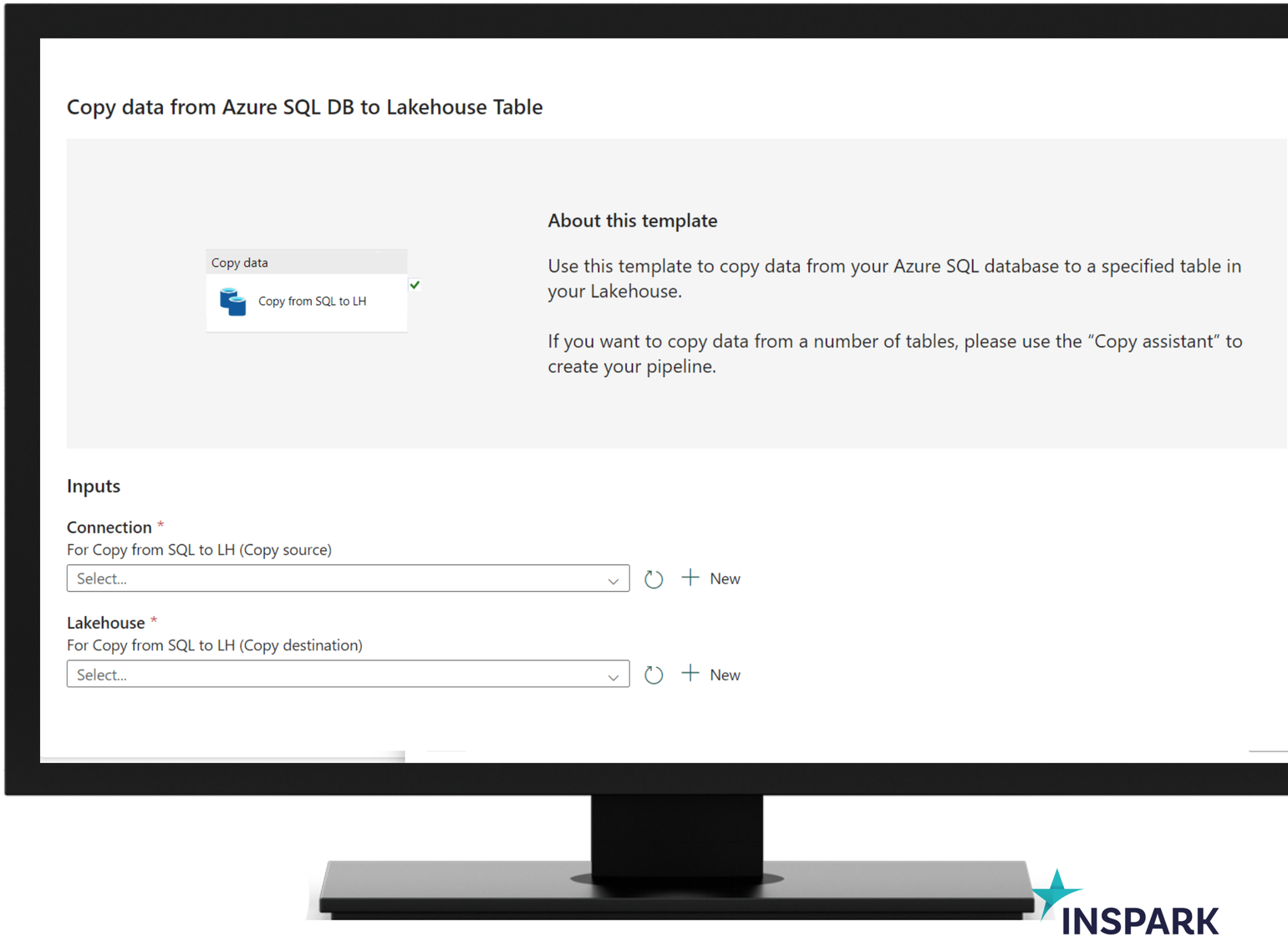



```
[
    {
        "source": {
            "table": "Department"
        },
        "destination": {
            "fileName": "HumanResourcesDepartment.parquet"
        }
    },
    {
        "source": {
            "table": "Employee"
        },
        "destination": {
            "fileName": "HumanResourcesEmployee.parquet"
        }
    },
```

INSPARK

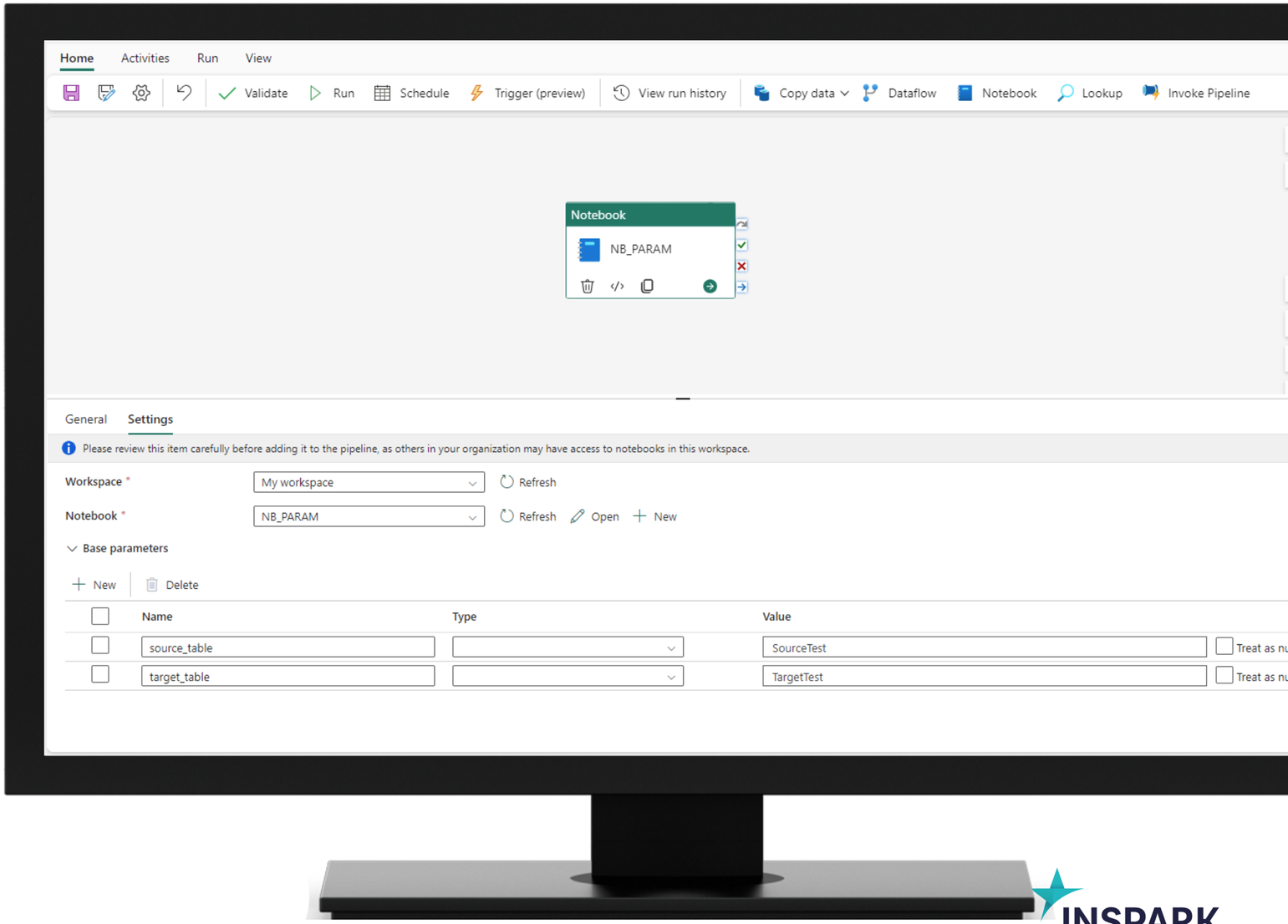
Templates

- Templates are pre-defined pipelines that allow you to get started quickly with Data Factory.
- These templates help to reduce development time by providing an easy way to create pipelines.
- Templates are available for common data integration scenarios.
- Templates can be customized to meet specific requirements



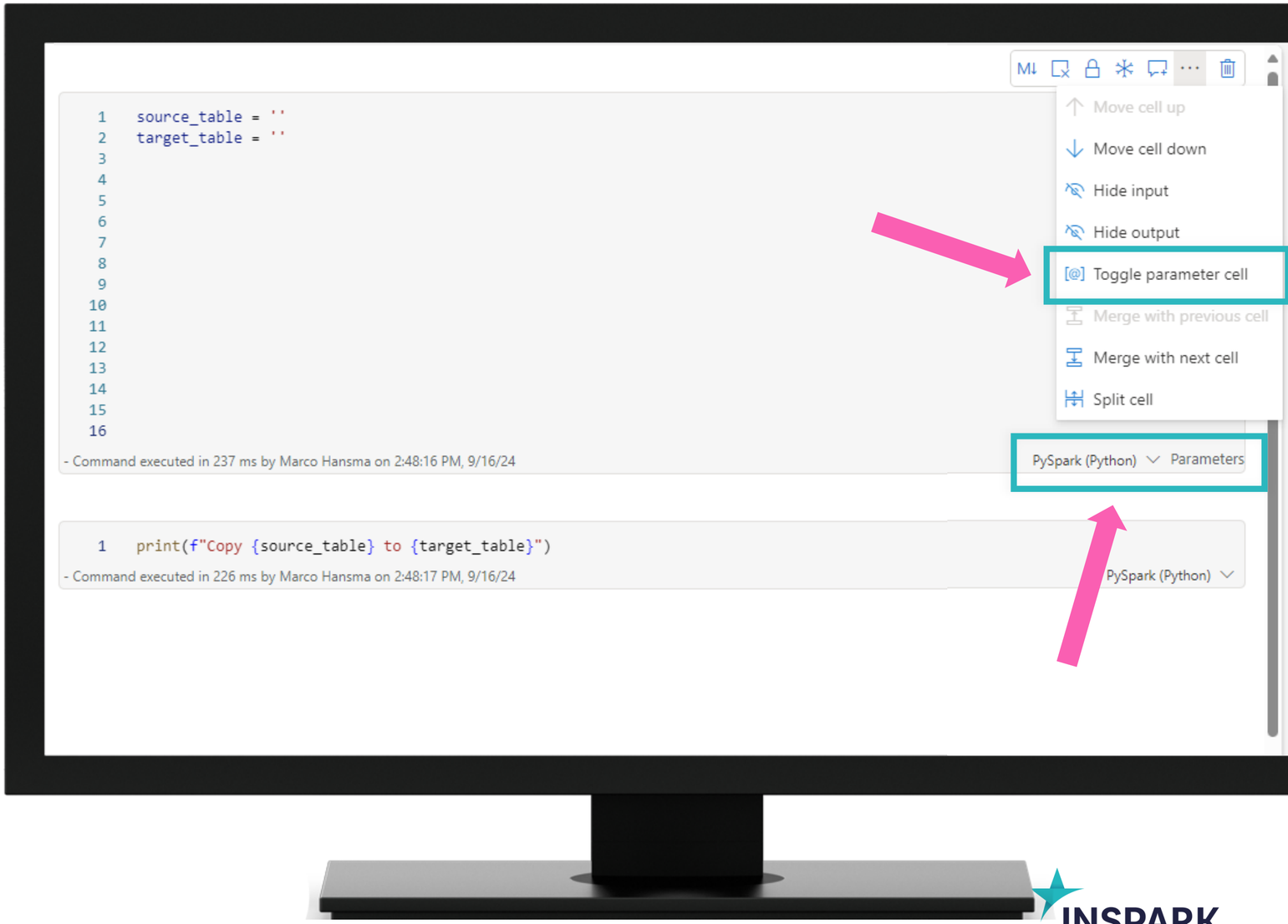
Blank canvas

- Everything build from scratch with no predefined options



Notebook Parameters

- Pass Parameters from Data Pipeline to Notebook
  - Toggle parameter cell

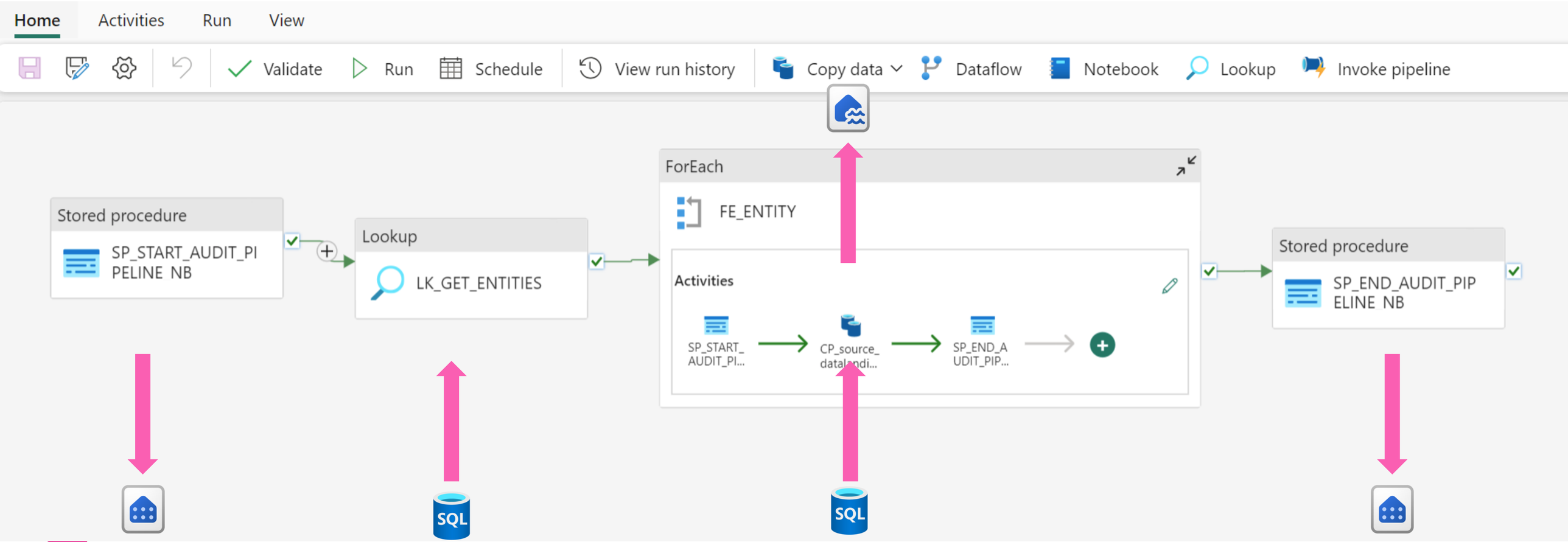


Innovate to accelerate

DEMO

Framework in Azure SQL

INSPARK


Home
Activities
Run
View
Validate
Run
Schedule
View run history
Copy data
Dataflow
Notebook
Lookup
Invoke pipeline
Stored procedure
SP_START_AUDIT_PIPELINE_NB
Lookup
LK_GET_ENTITIES
ForEach
FE_ENTITY
Activities
SP_START_AUDIT_PI...
CP_source_datalandi...
SP_END_AUDIT_PIP...
Stored procedure
SP_END_AUDIT_PIPELINE_NB
SQL
SQL

INSPARK

Framework

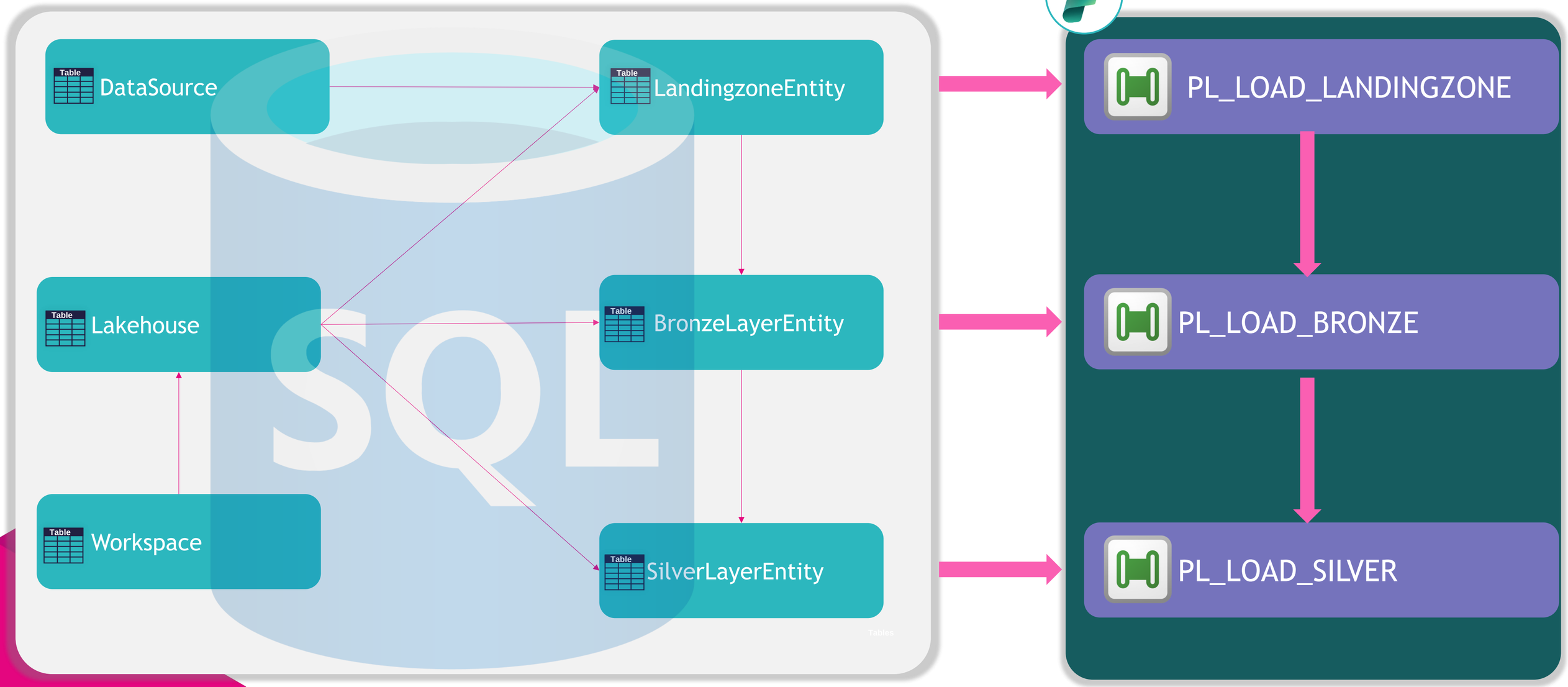
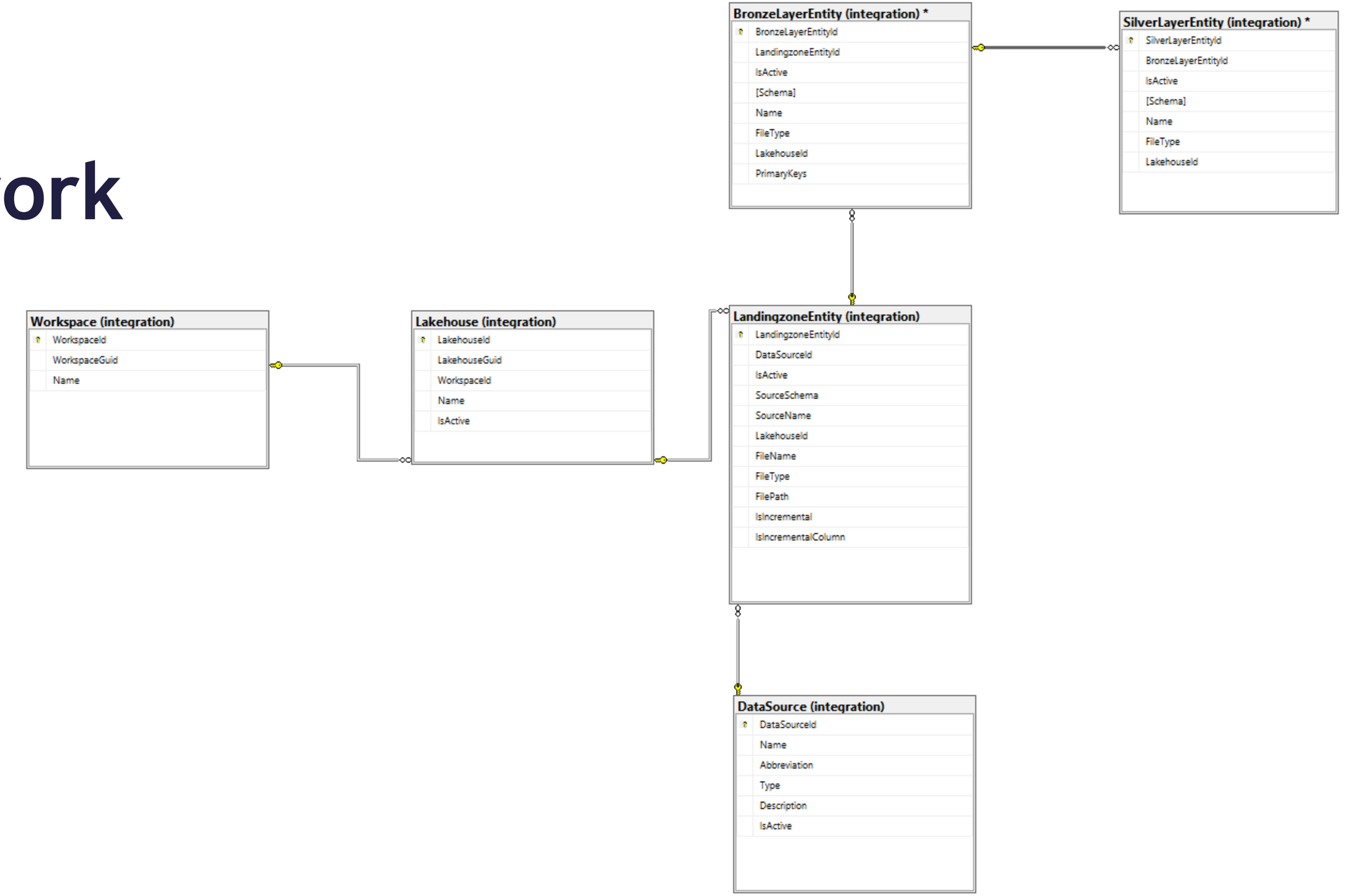


Table
DataSource
Table
LandingzoneEntity
Table
Lakehouse
Table
BronzeLayerEntity
Table
Workspace
Table
SilverLayerEntity
Tables
PL_LOAD_LANDINGZONE
PL_LOAD_BRONZE
PL_LOAD_SILVER

Framework


BronzeLayerEntity (integration) *
BronzeLayerEntityId
LandingzoneEntityId
IsActive
[Schema]
Name
FileType
Lakehouseld
PrimaryKeys
SilverLayerEntity (integration) *
SilverLayerEntityId
BronzeLayerEntityId
IsActive
[Schema]
Name
FileType
Lakehouseld
Workspace (integration)
Workspaceld
WorkspaceGuid
Name
Lakehouse (integration)
Lakehouseld
LakehouseGuid
Workspaceld
Name
IsActive
LandingzoneEntity (integration)
LandingzoneEntityId
DataSourceId
IsActive
SourceSchema
SourceName
Lakehouseld
FileName
FileType
FilePath
IsIncremental
IsIncrementalColumn
DataSource (integration)
DataSourceId
Name
Abbreviation
Type
Description
IsActive

Who is using a Medallion architecture?

‘Uniform data architecture’
From data “Spaghetti to Lasagna”

‘Data processing in different stages’

Medallion Architecture

Stages

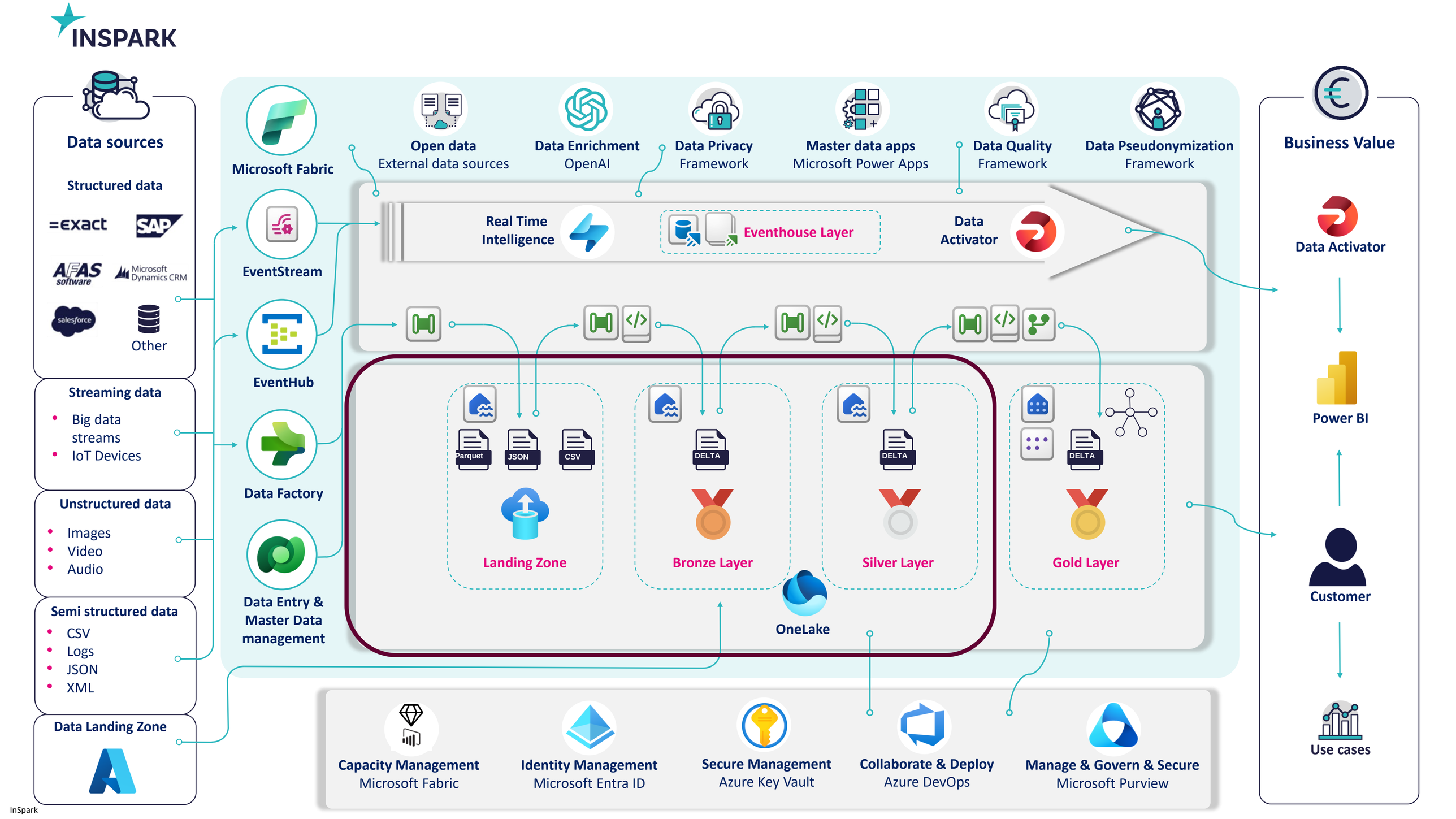
INSPARK

'Data processing in different stages'

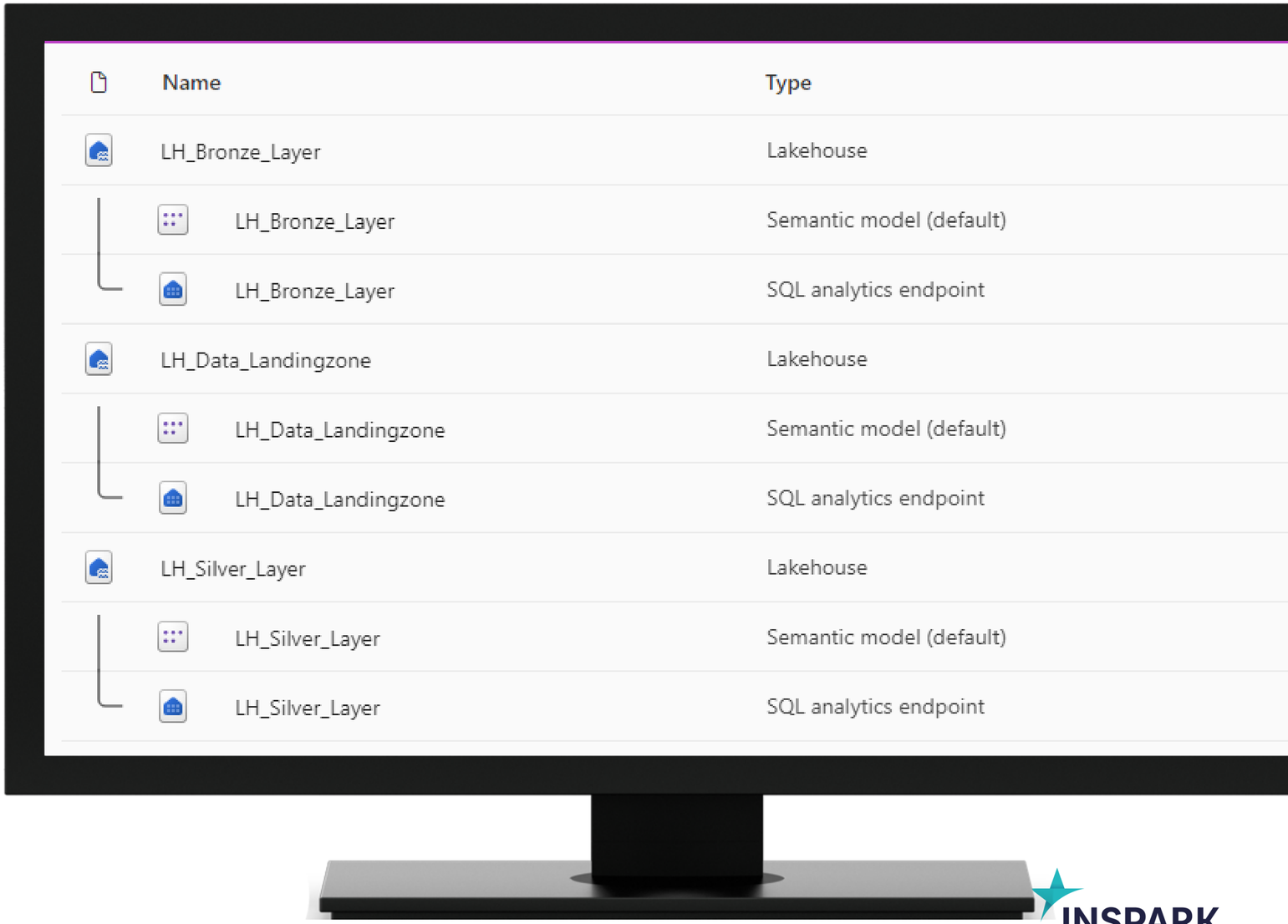
Medallion Architecture

Stage:	Definition:	Filetype:	Files/Tables:	Fabric:
Gold layer	• Dimensions & Facts (Star Schema) • Historical Analysis • Business rules • Documentation • Aggregated data • Logical table names	Delta	Tables	SPROC, View
Silver layer	• Historical Data (Type 1 or 2) • Data quality rules • Data Cleansing • Validated data • No business model/data	Delta	Tables	
Bronze layer	• Deduplicate data • Add datatypes • Data can be inconsistent • Mostly a copy of the source • Schema	Delta	Tables	
Landing zone	• Structured data • Unstructured data • Incremental loads • Data as is • Stored in Datetime folder structure • No Schema	JSON, CSV, Parquet	File, Tables	

INSPARK


INSPARK
Data sources
Structured data
Other
Streaming data
Big data streams
IoT Devices
Unstructured data
Images
Video
Audio
Semi structured data
CSV
Logs
JSON
XML
Data Landing Zone
Microsoft Fabric
EventStream
EventHub
Data Factory
Data Entry & Master Data management
Open data
External data sources
Data Enrichment
OpenAI
Data Privacy
Framework
Master data apps
Microsoft Power Apps
Data Quality
Framework
Data Pseudonymization
Framework
Real Time Intelligence
Eventhouse Layer
Data Activator
Parquet
JSON
CSV
DELTA
DELTA
DELTA
Landing Zone
Bronze Layer
Silver Layer
Gold Layer
OneLake
Capacity Management
Microsoft Fabric
Identity Management
Microsoft Entra ID
Secure Management
Azure Key Vault
Collaborate & Deploy
Azure DevOps
Manage & Govern & Secure
Microsoft Purview
Business Value
Data Activator
Power BI
Customer
Use cases

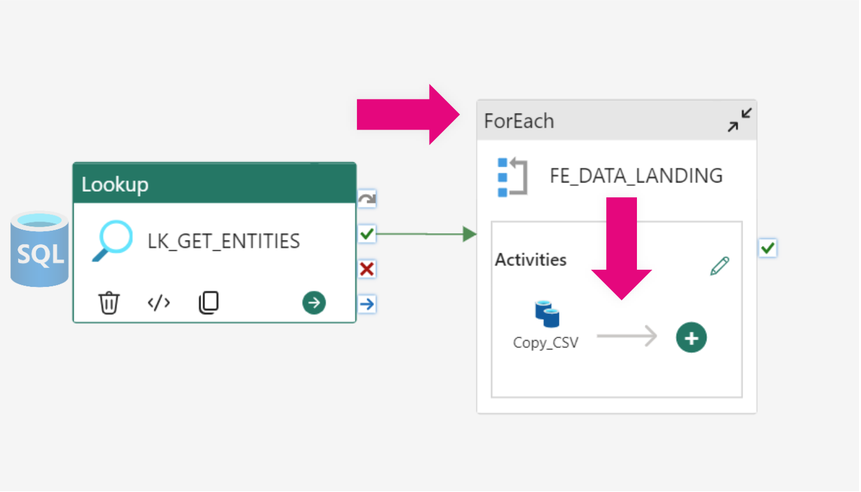
Lakehouse



INSPARK

Copy Activity Parameters

- Pass Parameters from Pipeline to Copy activity
- Use Parameters from For Each Activity



Preview data

ceName	DataSourceAbbreviation	DataSourceType	IsActive	SourceSchema	SourceName	TargetFilePath	TargetFileName	TargetFileType
uwdvlmdl01	ADLS	ADLS	true		customers.csv	demo/2024/09/10	customers_2024091015.csv	csv
uwdvlmdl01	ADLS	ADLS	true		organizations.csv	demo/2024/09/10	organizations_2024091015.csv	csv
uwdvlmdl01	ADLS	ADLS	true		people.csv	demo/2024/09/10	people_2024091015.csv	csv

General | Source | **Destination** | Mapping | Settings

Connection * : @item().TargetLakehouseGuid

Connection type * : Lakehouse

Workspace ID : @item().WorkspaceGuid

Root folder : Tables / Files

File path : @item().TargetFilePath / @item().TargetFileName Browse

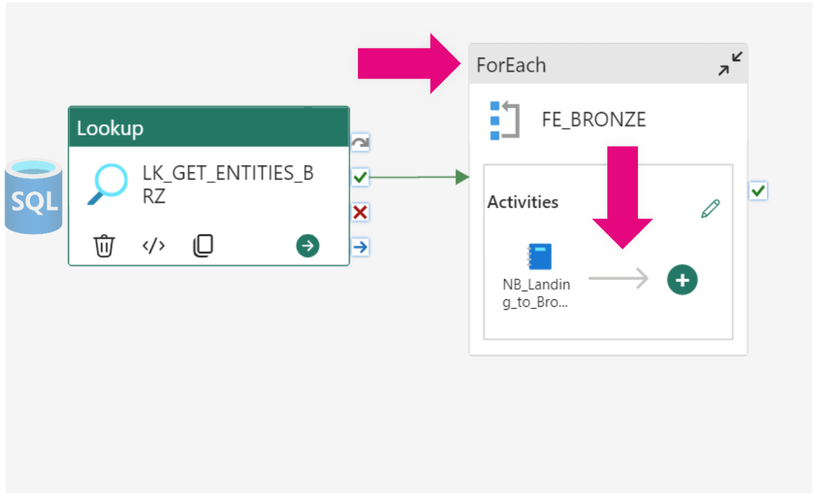
File format * : Binary Settings

Advanced

INSPARK

Notebook Parameters

- Pass Parameters from Data Pipeline to Notebook
  - Toggle parameter cell



Preview data

	EntityId	SourceFilePath	SourceFileName	SourceFileType	TargetSchema	TargetName	TargetWorkspaceId	SourceWorkspaceId	TargetLakehouseId	SourceLakehouseId	TargetLakehouseName	SourceLakehouseNa
1	1	demo/2024/09/10	customers_2024091017.csv	csv	demo	Customers	586fc19d-fa6a-4cb1-9ca3-e518a524f5da	586fc19d-fa6a-4cb1-9ca3-e518a524f5da	e130dba7-c8c3-438a-85ad-2cd4c9d59a09	009058ac-b71c-4774-a8ee-7c7d945c3972	LH_Bronze_Layer	LH_Data_Landingzor
2	3	demo/2024/09/10	organizations_2024091017.csv	csv	demo	Organizations	586fc19d-fa6a-4cb1-9ca3-e518a524f5da	586fc19d-fa6a-4cb1-9ca3-e518a524f5da	e130dba7-c8c3-438a-85ad-2cd4c9d59a09	009058ac-b71c-4774-a8ee-7c7d945c3972	LH_Bronze_Layer	LH_Data_Landingzor
3	4	demo/2024/09/10	people_2024091017.csv	csv	demo	People	586fc19d-fa6a-4cb1-9ca3-e518a524f5da	586fc19d-fa6a-4cb1-9ca3-e518a524f5da	e130dba7-c8c3-438a-85ad-2cd4c9d59a09	009058ac-b71c-4774-a8ee-7c7d945c3972	LH_Bronze_Layer	LH_Data_Landingzor

Notebook * NB_LANDING_BRONZE Refresh Open New

Base parameters

New Delete

Name	Type	Value
SourceLakehouse	String	@item().SourceLakehouseId
source_file_path	String	@item().SourceFilePath
source_file_name	String	@item().SourceFileName
PrimaryKeys	String	@item().PrimaryKeys
TargetLakehouse	String	@item().TargetLakehouseId
target_schema	String	@item().TargetSchema
target_name	String	@item().TargetName
SourceWorkspace	String	@item().SourceWorkspaceId
TargetWorkspace	String	@item().TargetWorkspaceId
source_file_type	String	@item().SourceFileType

INSPARK

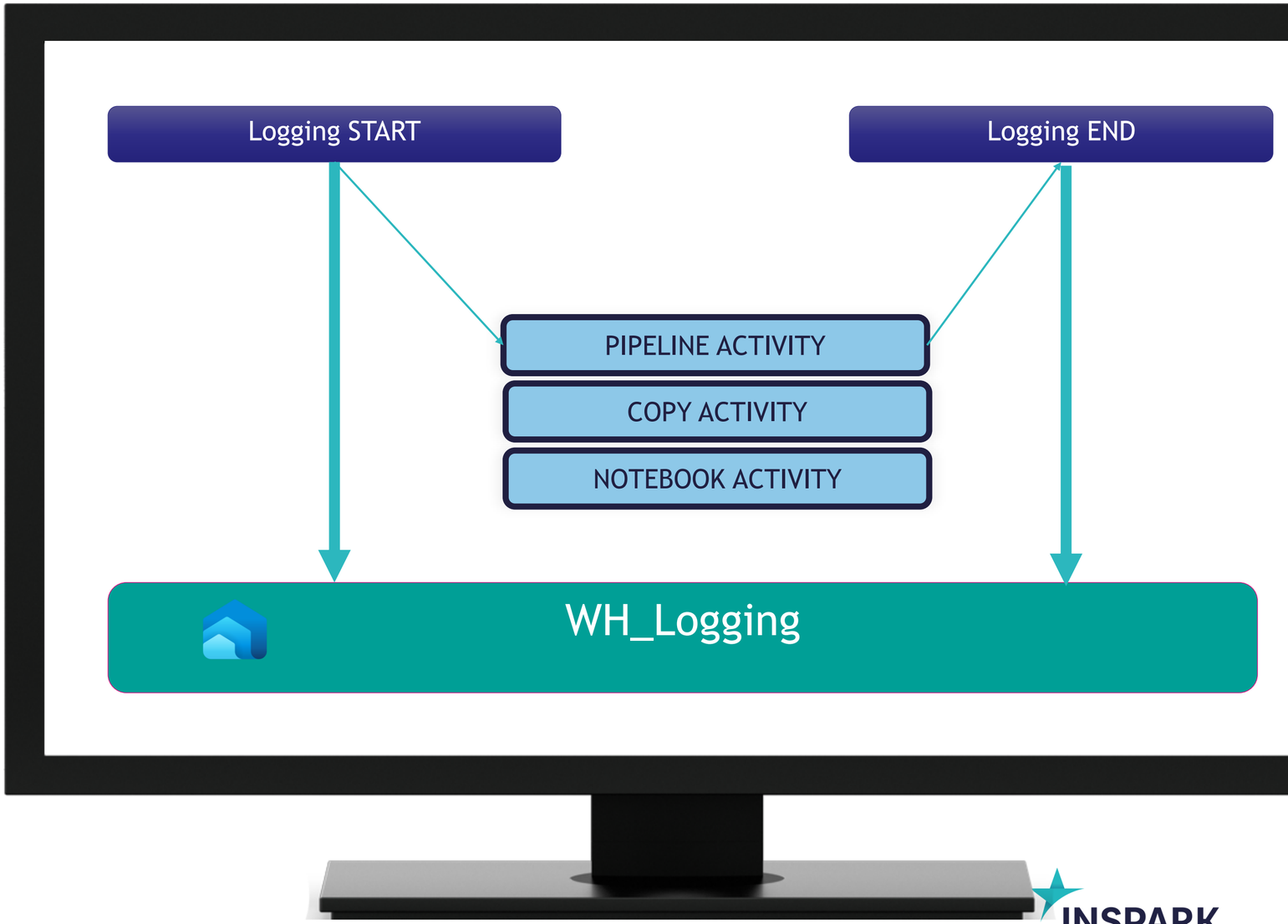
Innovate to accelerate

DEMO

Logging

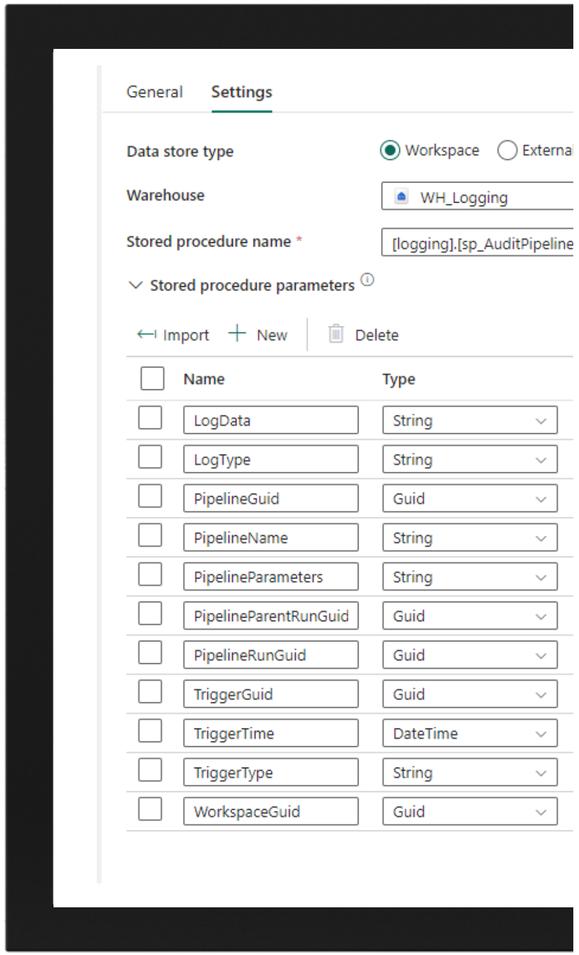
Logging

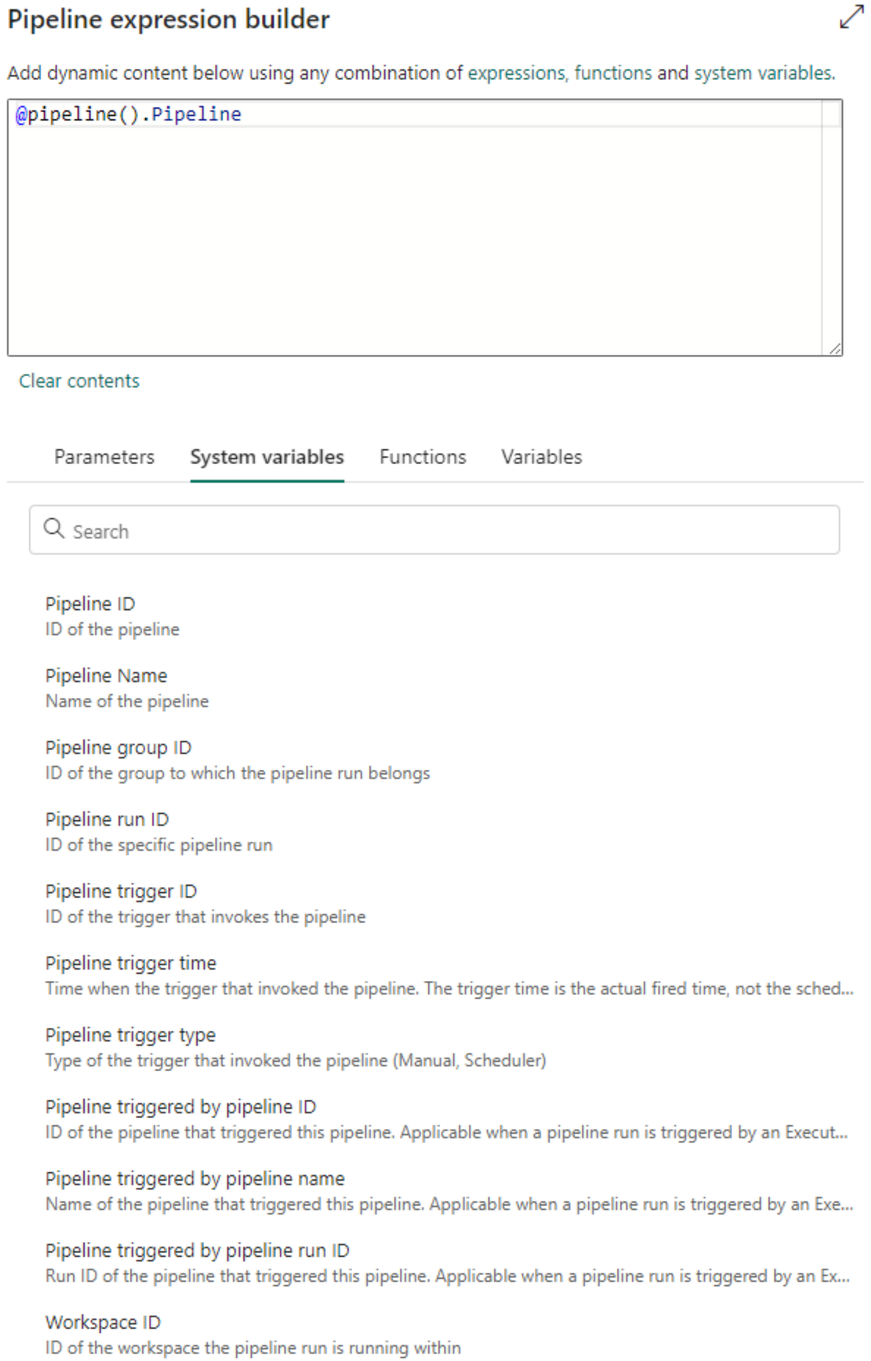
- Log Start and End Time of records
- Log Extracted Records
- Log Execution Failure



Logging

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





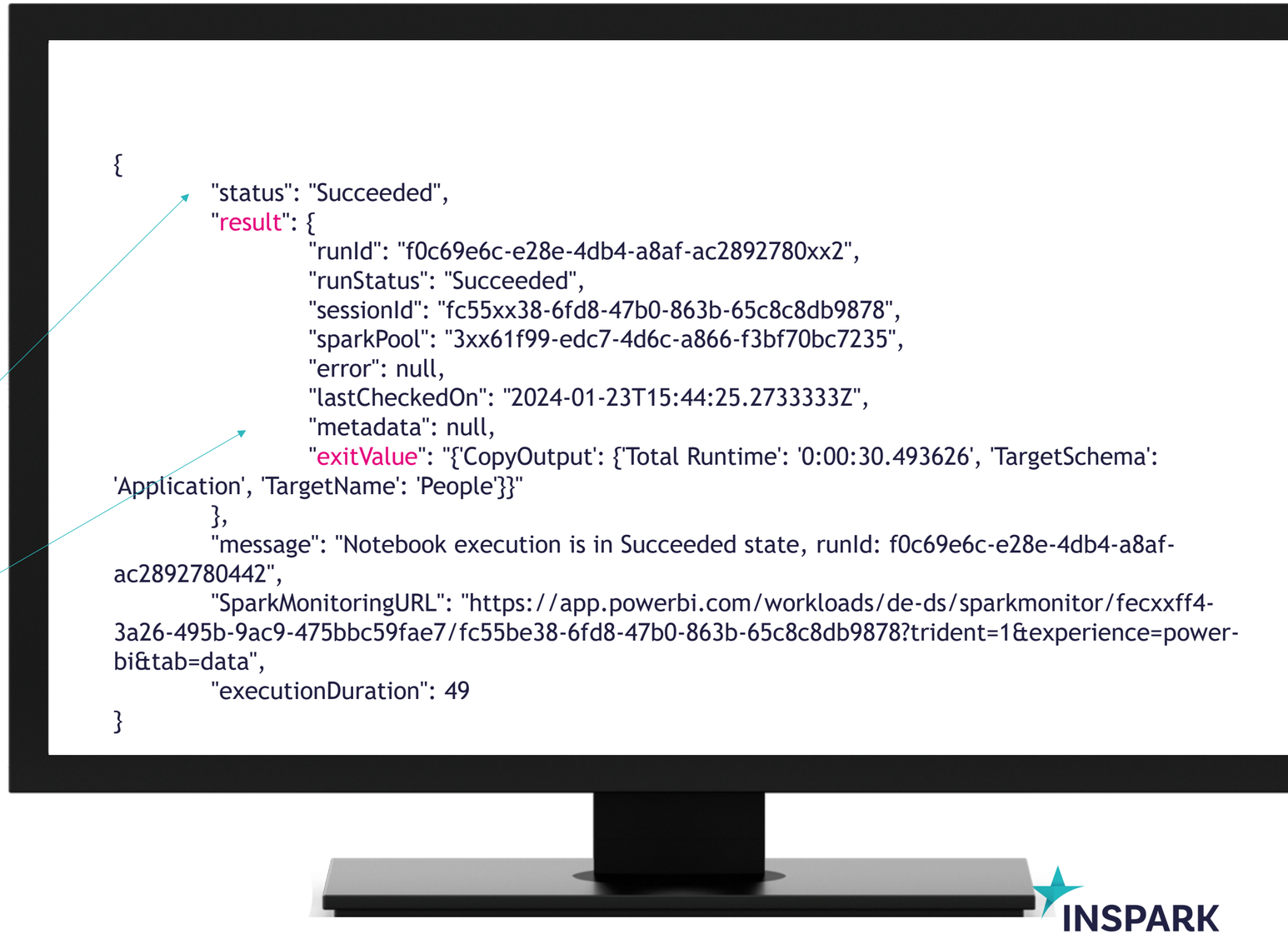
Logging

- Add Information about pipelines
- Adding System Variables
- Add Information about Notebooks

Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variables.

```
{
    "Action" : "End",
    @{activity('NB_Landing_to_Bronze').output.result.exitValue}
}
```



Innovate to accelerate

DEMO

Challenges

- Out of the box fast and easy, less flexible
- Parameterize of connections with Azure Key Vault (Like ADF/Synapse with Linked Services)
- Schedule can’t be Parameterized like in ADF/Synapse
- Build in retry to Notebook Activity

INSPARK

Key Learnings and Best Practices

Metadata-driven approach is a best practice for managing data pipelines

Medallion Lakehouse architecture is a proven framework for implementing metadata-driven approach

Start small and gradually expand the metadata-driven approach across the organization

Session Feedback 💖

https://bit.ly/dMC2024_SessionFeedback

INSPARK

Thank you, partners

element 61
experience & expertise

Cloubis

Microsoft

MICROPOLE BELUX

PLAINSIGHT

Adapt and Enable

Tabular Editor

delaware

datashift

U2U

cegeka

EpicData.

LACO

tillit data shapers

ilytix

bmatix
Act informed

randstad digital

KOHERA

MONIN
Database Managed Services

inetum realdolmen
Positive digital flow

TriFinance
BEYOND ADVISORY

@erwindekreuk

linkedin.com/in/erwindekreuk

erwindekreuk.com

github.com/edkreuk

https://sessionize.com/erwin-de-kreuk/

linkedin.com/in/marcohansma

Microsoft® MVP Most Valuable Professional

INSPARK