12/19/
2019

MOVING DATA AROUND IN AZURE

moving data around in Azure

In the Microsoft Data platform ecosystem in Azure, there are several possibilities to move data around. The choice can be overwhelming and it's not always clear which option should be used.

In this session, we aim to give you an overview of a couple of products in Azure for data pipelines: Azure Data Factory, Logic Apps, Integration Services et cetera. At the end, you'll have a better understanding of the various alternatives in Azure.

Koen Verbeeck

contact information

koen dot verbeeck at ae dot be

@Ko_Ver

LinkedIn

https://sqlkover.com

MVP Microsoft® Most Valuable Professional

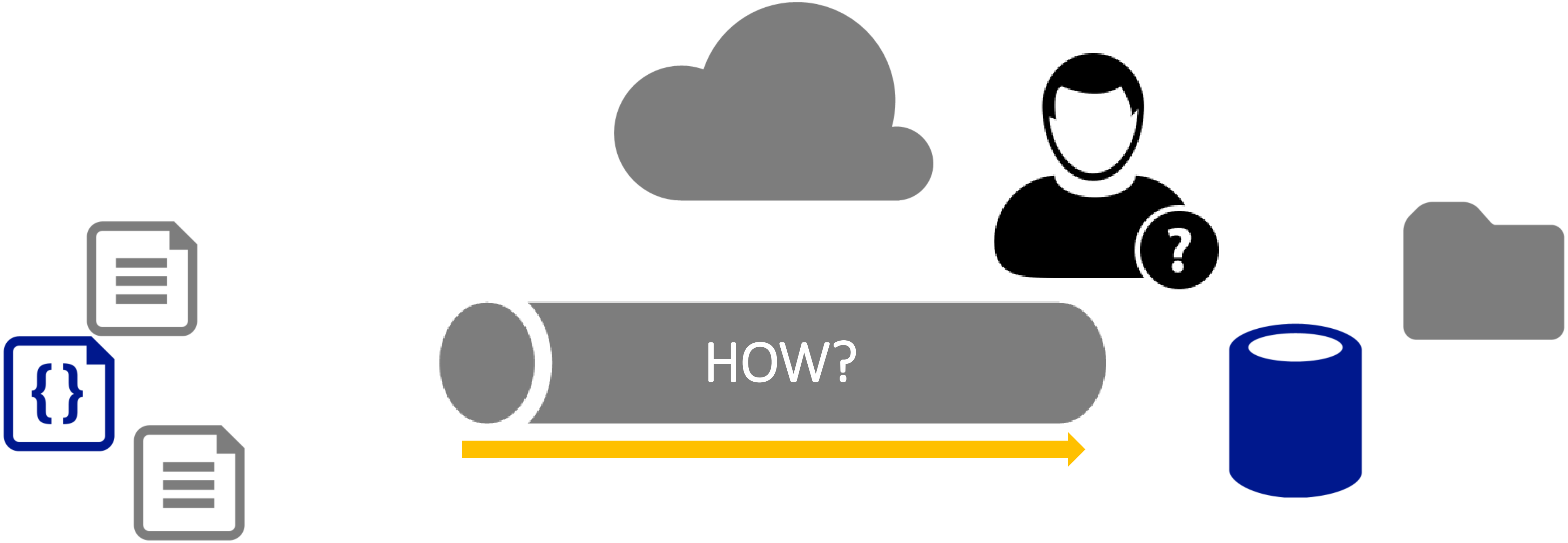
ae® architects for business & ict

MSSQLTips.com

Data Science | MICROSOFT PROFESSIONAL PROGRAM

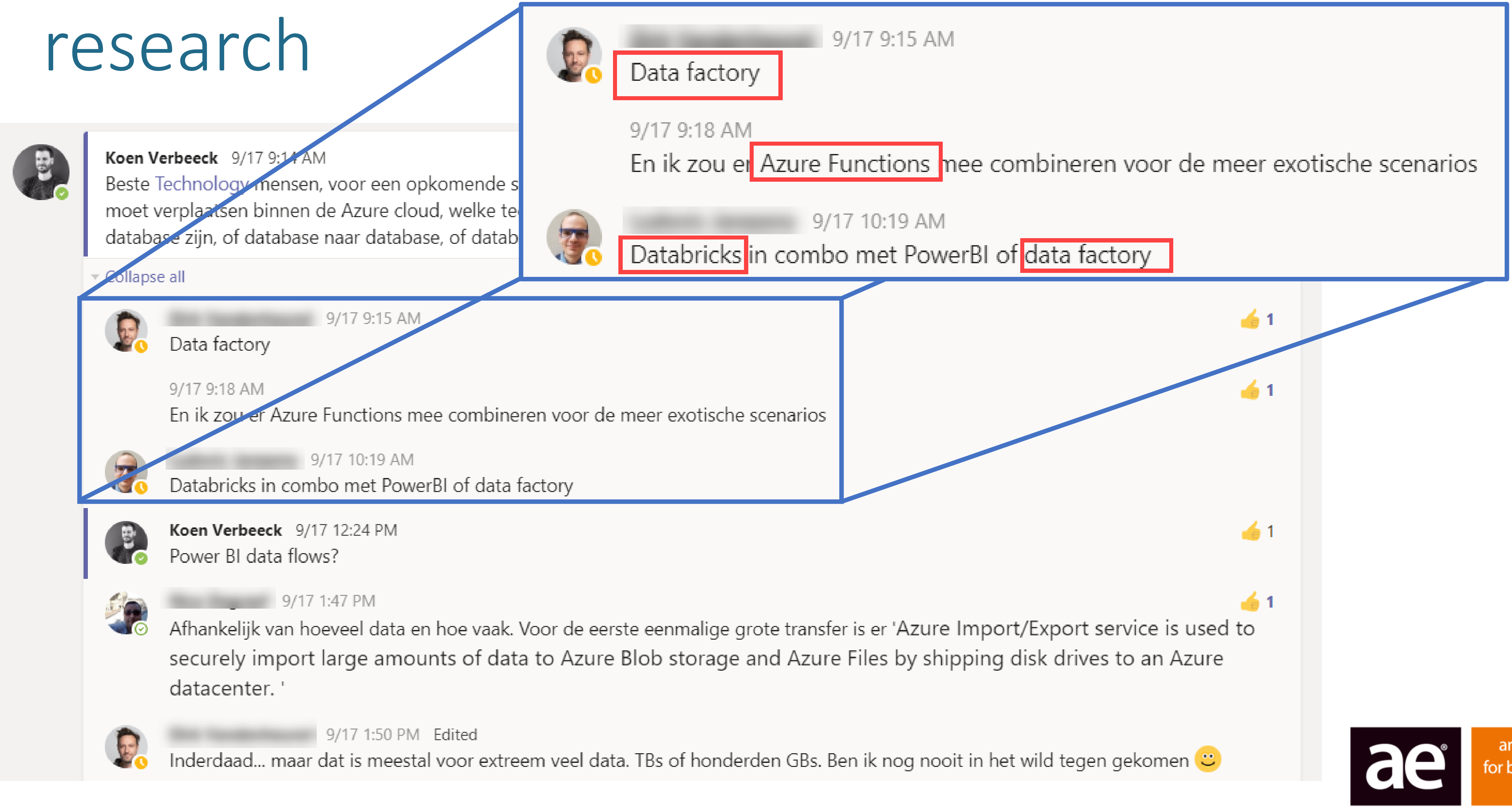
MCSE

the goal of this session



architects for business & ict

research



Koen Verbeeck 9/17 9:14 AM
Beste Technology mensen, voor een opkomende s
moet verplaatsen binnen de Azure cloud, welke te
database zijn, of database naar database, of datab

Collapse all

[redacted] 9/17 9:15 AM
Data factory
👍 1

9/17 9:18 AM
En ik zou er Azure Functions mee combineren voor de meer exotische scenarios
👍 1

[redacted] 9/17 10:19 AM
Databricks in combo met PowerBI of data factory

Koen Verbeeck 9/17 12:24 PM
Power BI data flows?
👍 1

[redacted] 9/17 1:47 PM
Afhankelijk van hoeveel data en hoe vaak. Voor de eerste eenmalige grote transfer is er 'Azure Import/Export service is used to securely import large amounts of data to Azure Blob storage and Azure Files by shipping disk drives to an Azure datacenter. '
👍 1

[redacted] 9/17 1:50 PM Edited
Inderdaad... maar dat is meestal voor extreem veel data. TBs of honderden GBs. Ben ik nog nooit in het wild tegen gekomen 🙂

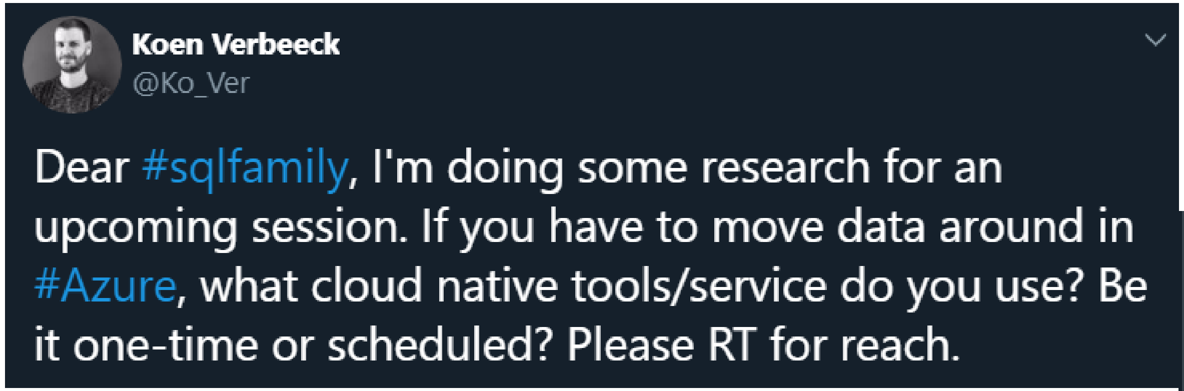
[redacted] 9/17 9:15 AM
Data factory

9/17 9:18 AM
En ik zou er Azure Functions mee combineren voor de meer exotische scenarios

[redacted] 9/17 10:19 AM
Databricks in combo met PowerBI of data factory

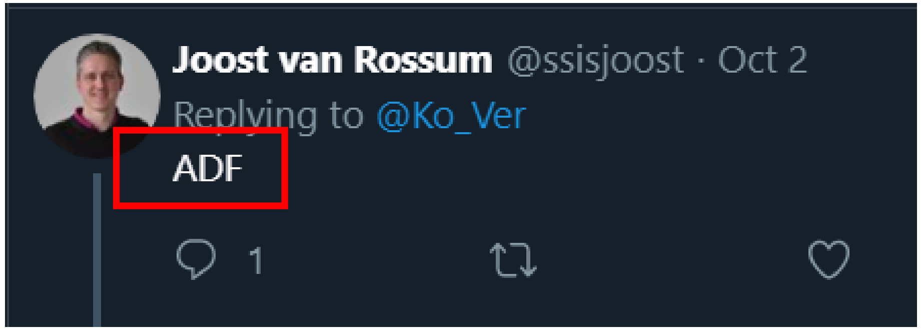
ae architects for business & ict

research



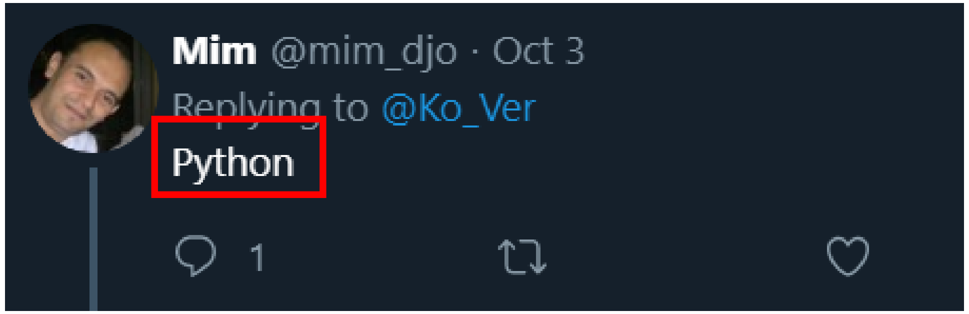
Koen Verbeeck
@Ko_Ver

Dear #sqlfamily, I'm doing some research for an upcoming session. If you have to move data around in #Azure, what cloud native tools/service do you use? Be it one-time or scheduled? Please RT for reach.



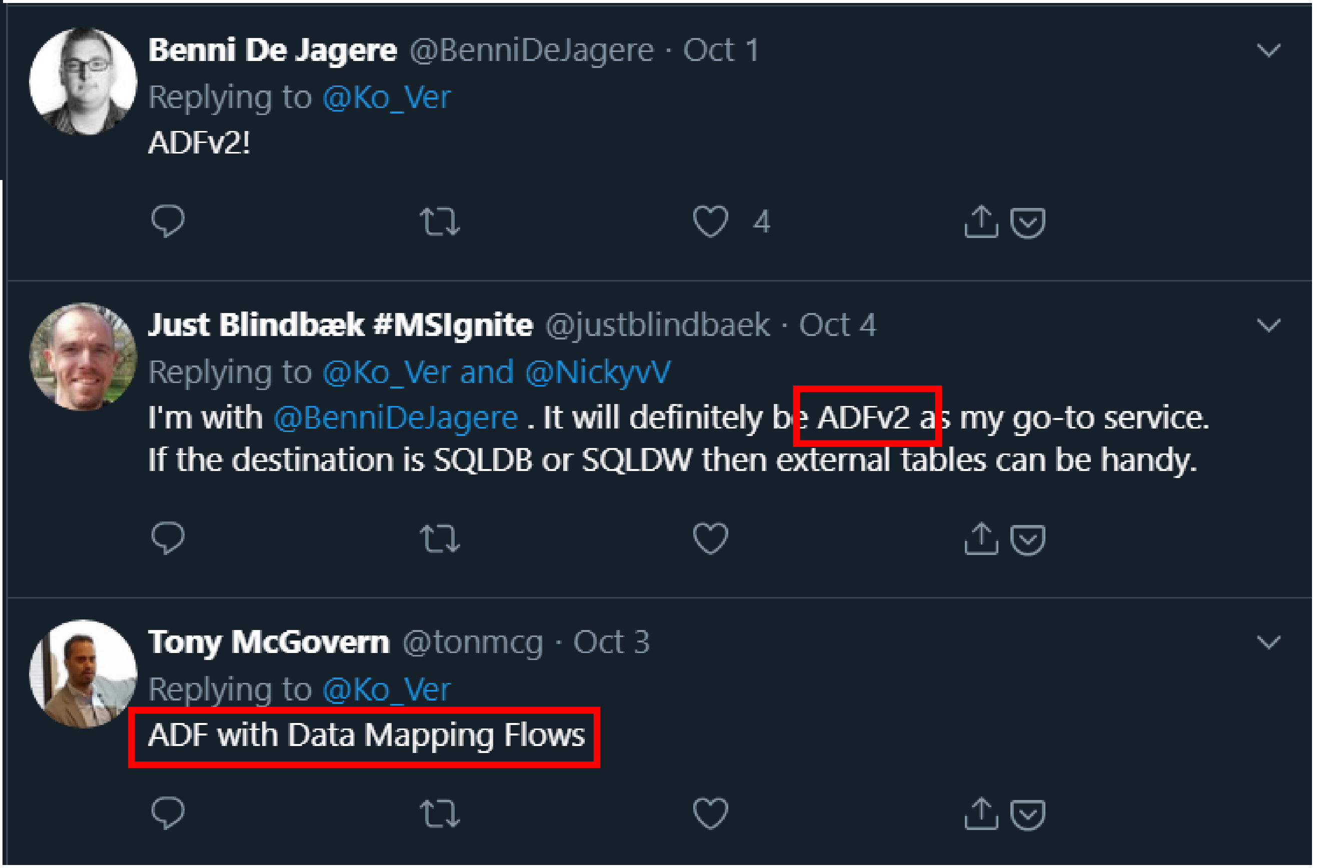
Joost van Rossum @ssisjoost · Oct 2
Replying to @Ko_Ver
ADF

1



Mim @mim_djo · Oct 3
Replying to @Ko_Ver
Python

1



Benni De Jagere @BenniDeJagere · Oct 1
Replying to @Ko_Ver
ADFv2!

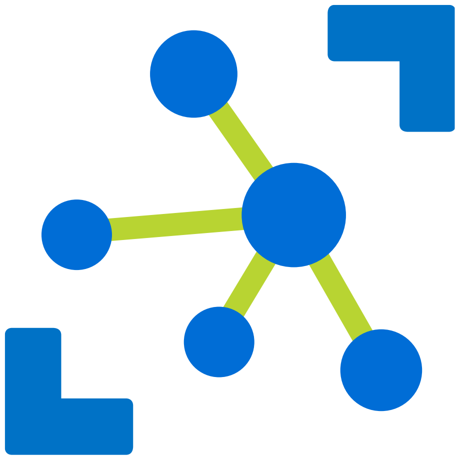
4

Just Blindbæk #MSIgnite @justblindbaek · Oct 4
Replying to @Ko_Ver and @NickyvV
I'm with @BenniDeJagere . It will definitely be ADFv2 as my go-to service. If the destination is SQLDB or SQLDW then external tables can be handy.

Tony McGovern @tonmcg · Oct 3
Replying to @Ko_Ver
ADF with Data Mapping Flows

what we’ll
not
be talking about

specific use cases

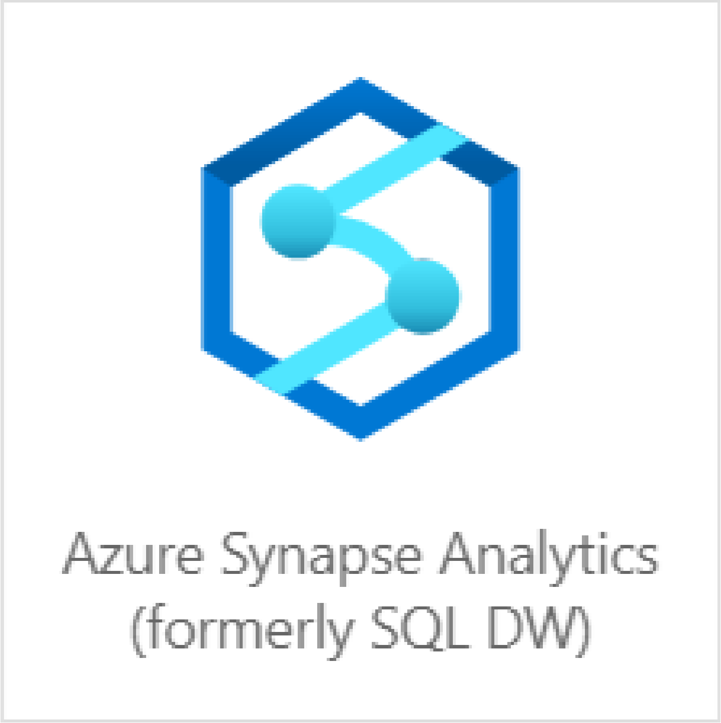


anything that runs on a VM

… or some sort of backup/migration service

ae architects for business & ict

also any service announced when I was making this slide deck



what we will be talking about

some Azure “data” services

writing code

automation

databricks

functions

not writing code

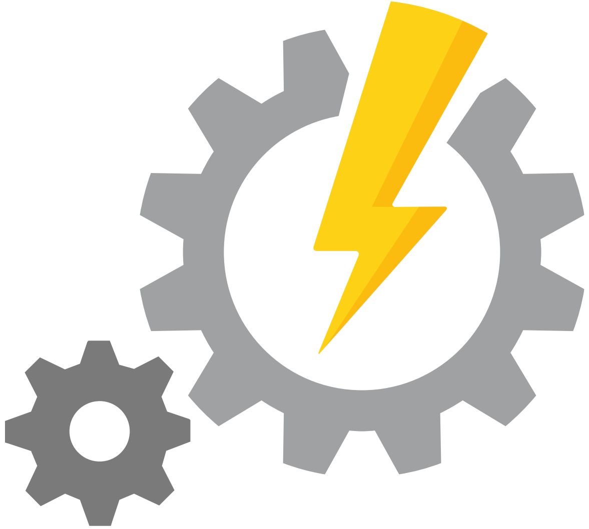
logic apps

data factory

azure automation

architects
for business
& ict

azure automation



aka runbooks

can run Python & PowerShell

no Visual Studio integration

better suited for automating administrative tasks

azure functions

architects
for business
& ict

azure functions

serverless compute

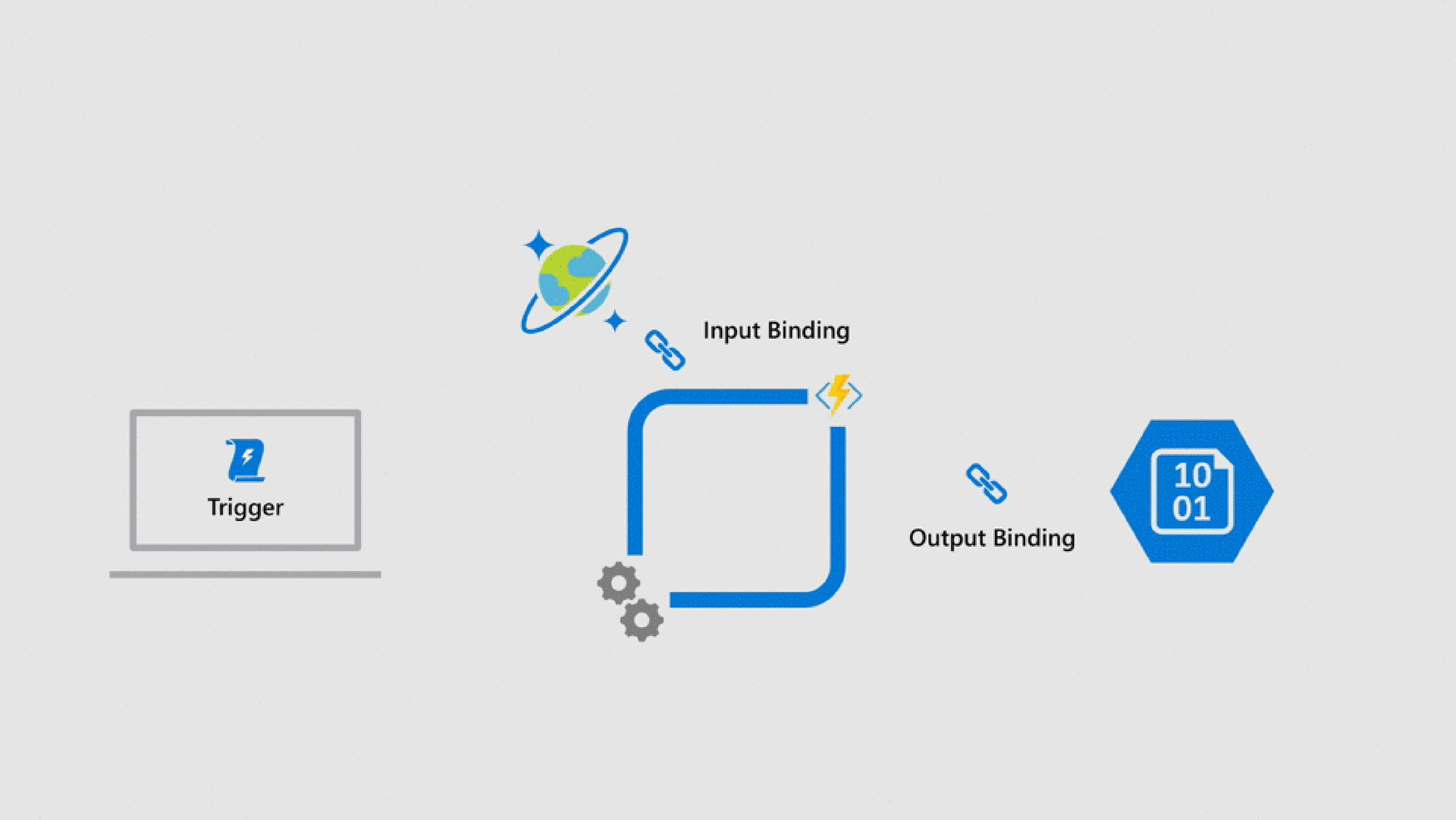
event-driven

Visual Studio integration

C#, PowerShell, Python, Batch, Bash, JavaScript, PHP or F#

source: https://azure.microsoft.com/en-us/services/functions/

azure functions

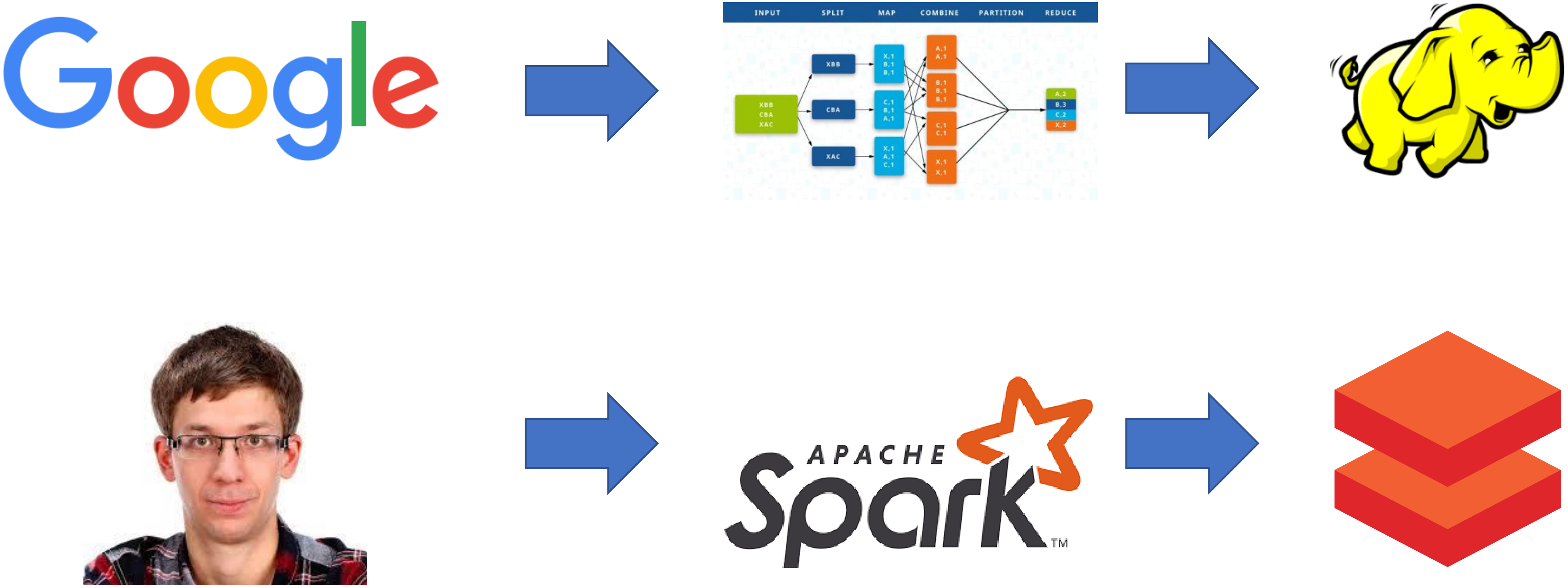


DEMO TIME

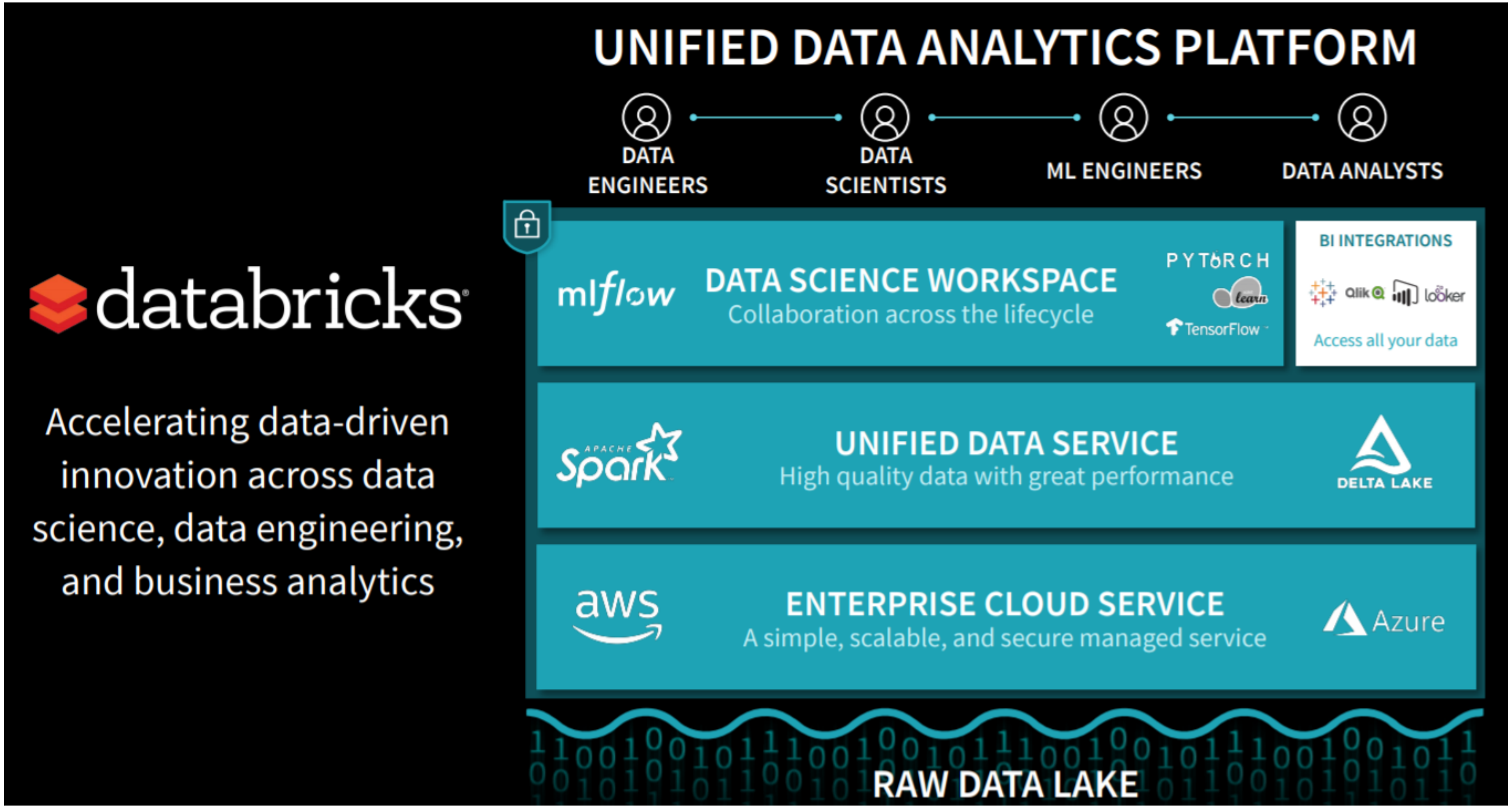
azure databricks

architects
for business
& ict

azure databricks



azure databricks



databricks

Accelerating data-driven innovation across data science, data engineering, and business analytics

UNIFIED DATA ANALYTICS PLATFORM

DATA ENGINEERS — DATA SCIENTISTS — ML ENGINEERS — DATA ANALYSTS

DATA SCIENCE WORKSPACE
Collaboration across the lifecycle

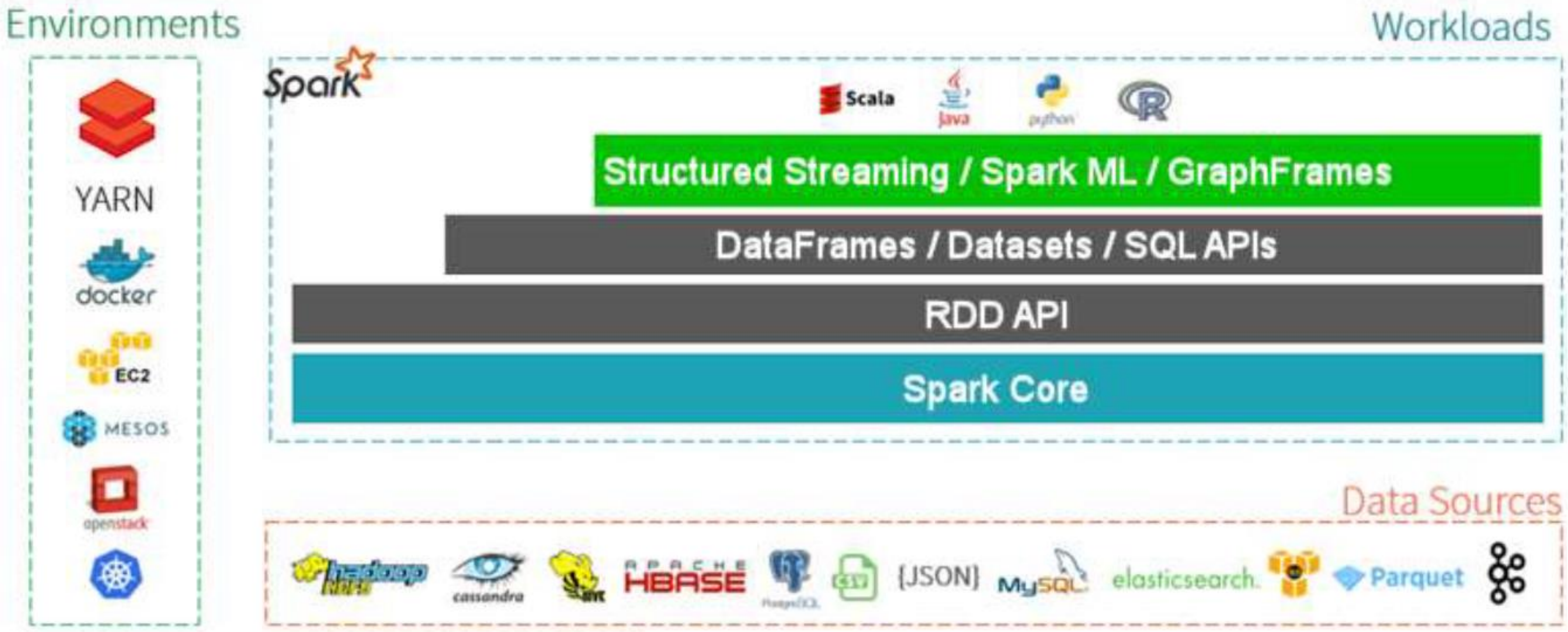
BI INTEGRATIONS
Access all your data

UNIFIED DATA SERVICE
High quality data with great performance

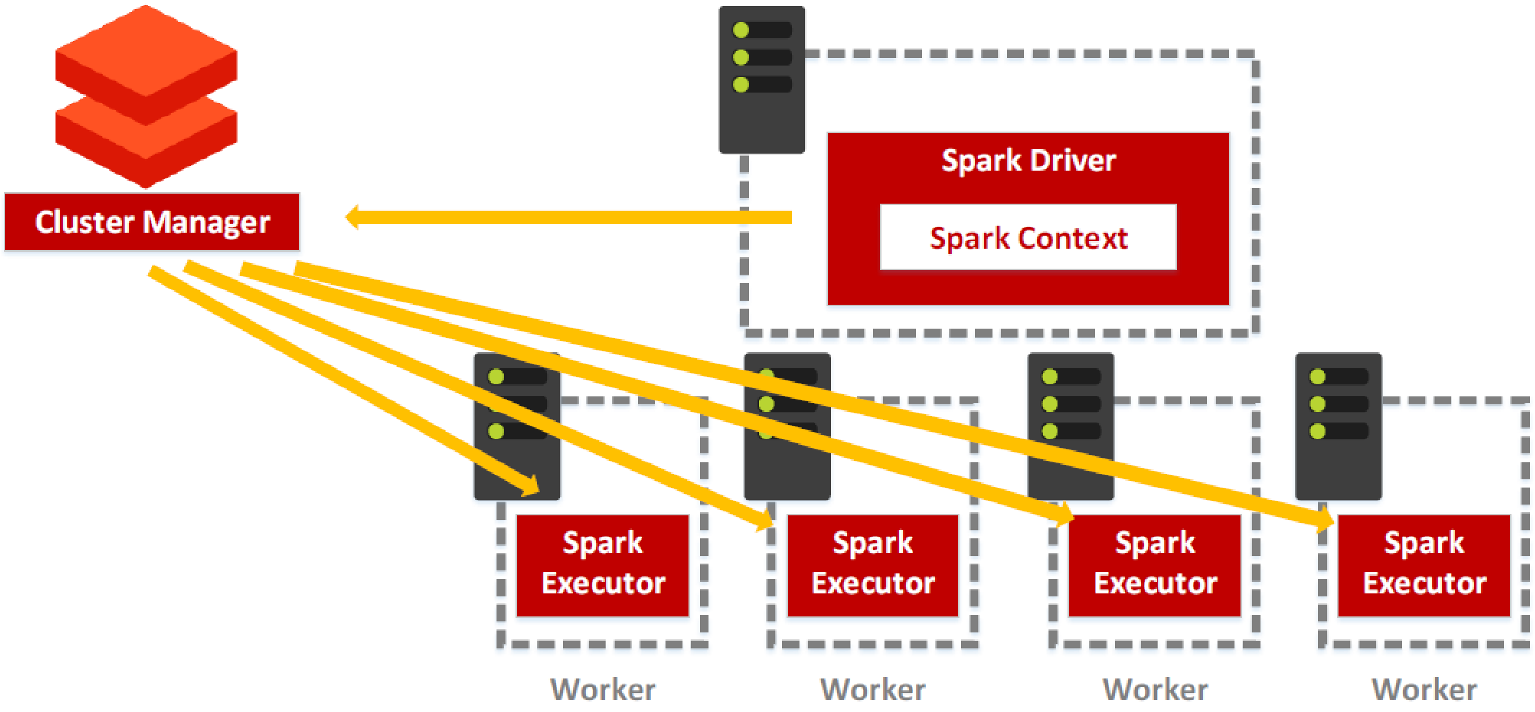
ENTERPRISE CLOUD SERVICE
A simple, scalable, and secure managed service

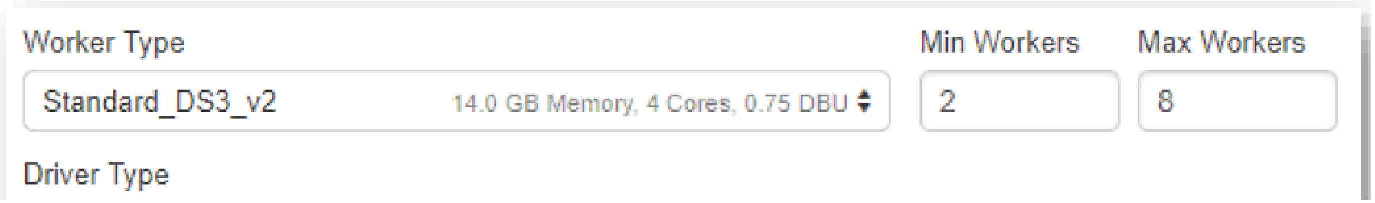
RAW DATA LAKE

azure databricks



azure databricks

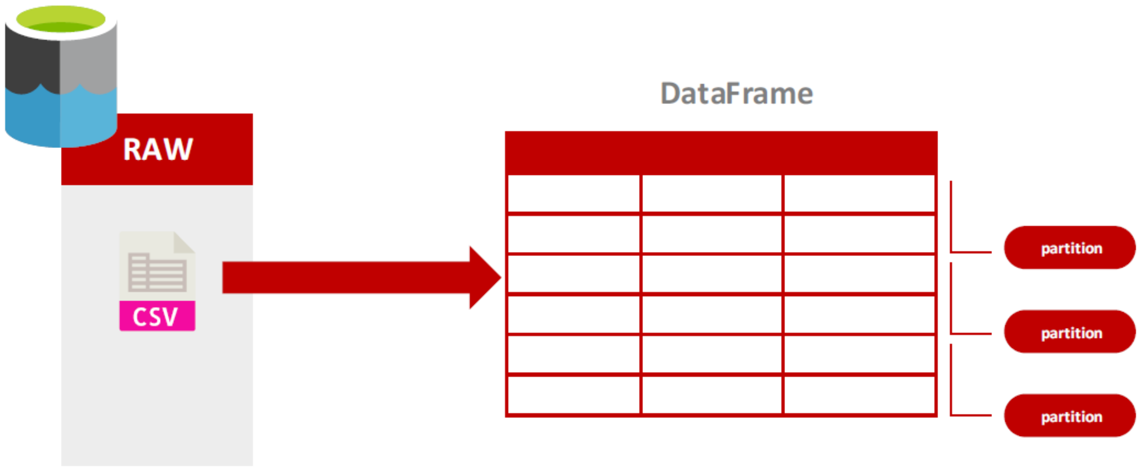


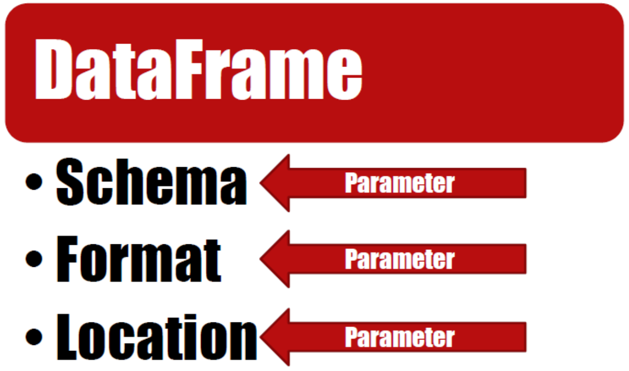


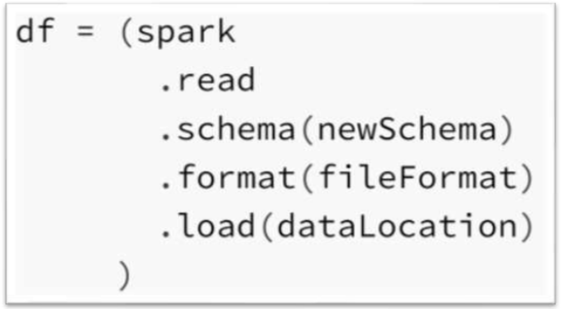
Worker Type	Min Workers	Max Workers
Standard_DS3_v2 14.0 GB Memory, 4 Cores, 0.75 DBU	2	8

Driver Type

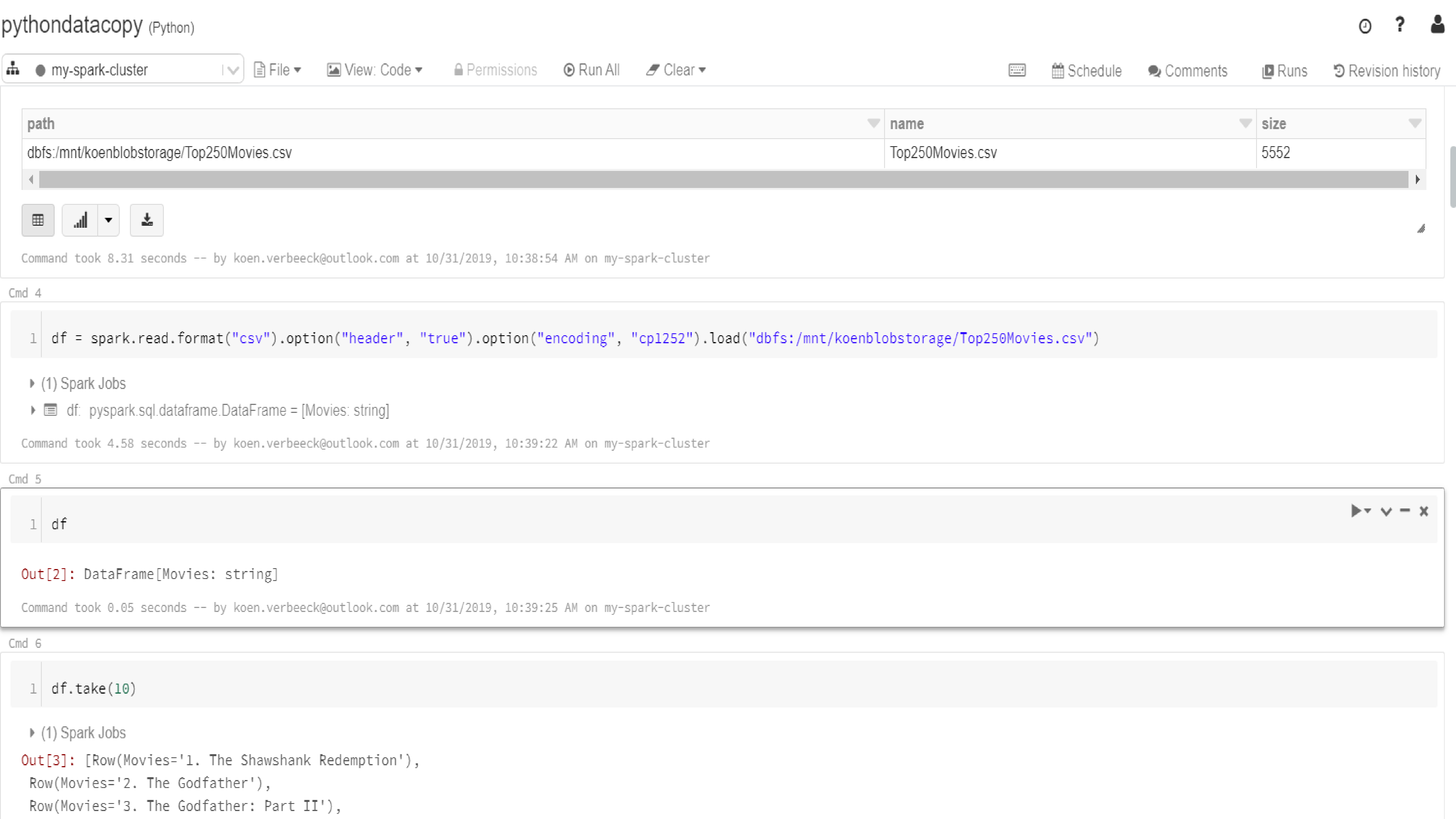
azure databricks







```
df = (spark
        .read
        .schema(newSchema)
        .format(fileFormat)
        .load(dataLocation)
    )
```



path	name	size
dbfs:/mnt/koenblobstorage/Top250Movies.csv	Top250Movies.csv	5552

```
Command took 8.31 seconds -- by koen.verbeeck@outlook.com at 10/31/2019, 10:38:54 AM on my-spark-cluster
```

Cmd 4

```
df = spark.read.format("csv").option("header", "true").option("encoding", "cp1252").load("dbfs:/mnt/koenblobstorage/Top250Movies.csv")
```

```
(1) Spark Jobs
df:  pyspark.sql.dataframe.DataFrame = [Movies: string]
Command took 4.58 seconds -- by koen.verbeeck@outlook.com at 10/31/2019, 10:39:22 AM on my-spark-cluster
```

Cmd 5

```
df
```

```
Out[2]: DataFrame[Movies: string]
Command took 0.05 seconds -- by koen.verbeeck@outlook.com at 10/31/2019, 10:39:25 AM on my-spark-cluster
```

Cmd 6

```
df.take(10)
```

```
(1) Spark Jobs
Out[3]: [Row(Movies='1. The Shawshank Redemption'),
 Row(Movies='2. The Godfather'),
 Row(Movies='3. The Godfather: Part II'),
```

DEMO TIME

architects
for business
& ict

azure logic apps

ae architects for business & ict

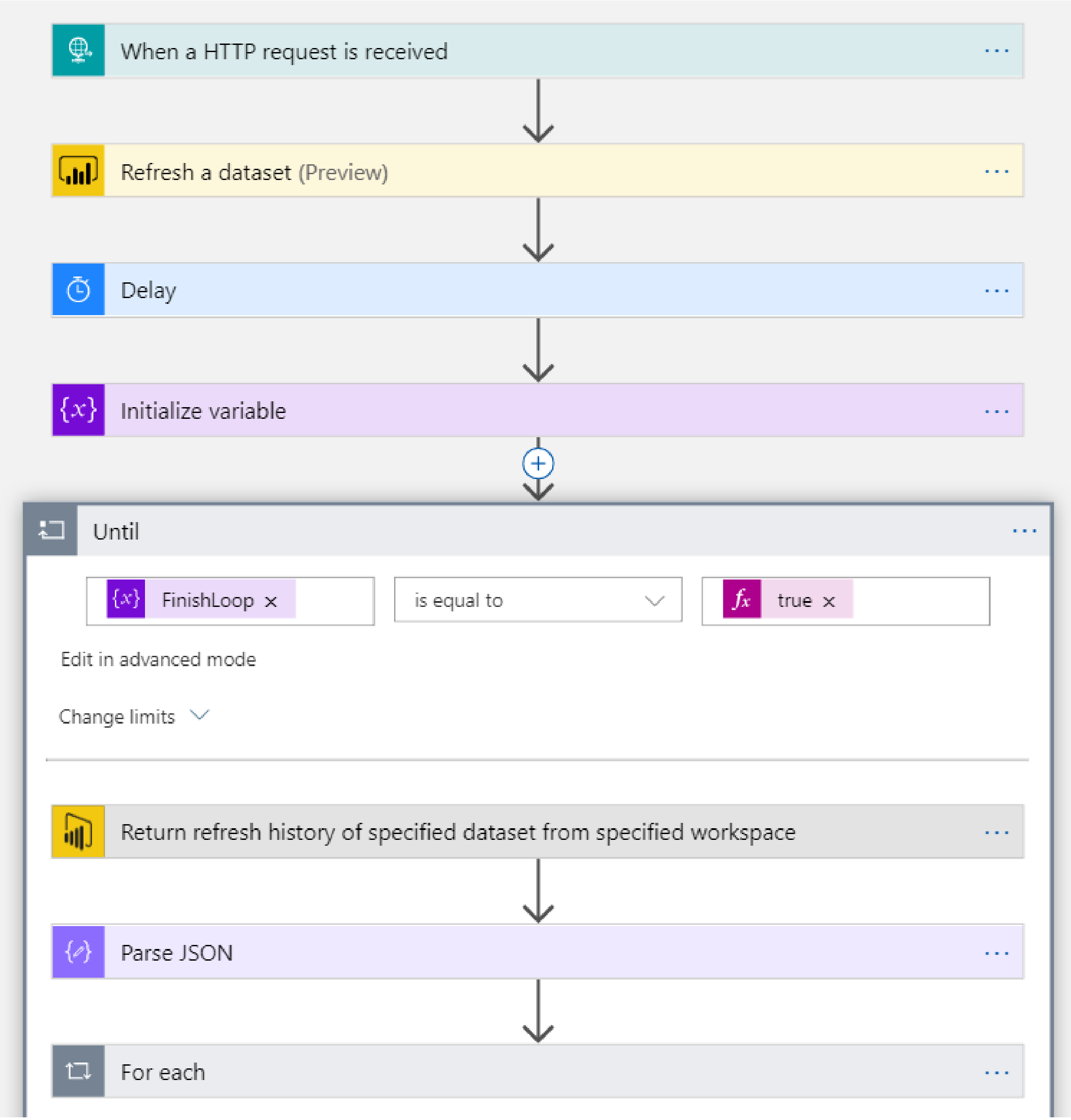
azure logic apps

“business workflows without coding”

lots of built-in connectors and actions

Azure version of ~~Microsoft Flow~~

Power Automate



When a HTTP request is received

Refresh a dataset (Preview)

Delay

Initialize variable

Until

FinishLoop × | is equal to | true ×

Edit in advanced mode

Change limits

Return refresh history of specified dataset from specified workspace

Parse JSON

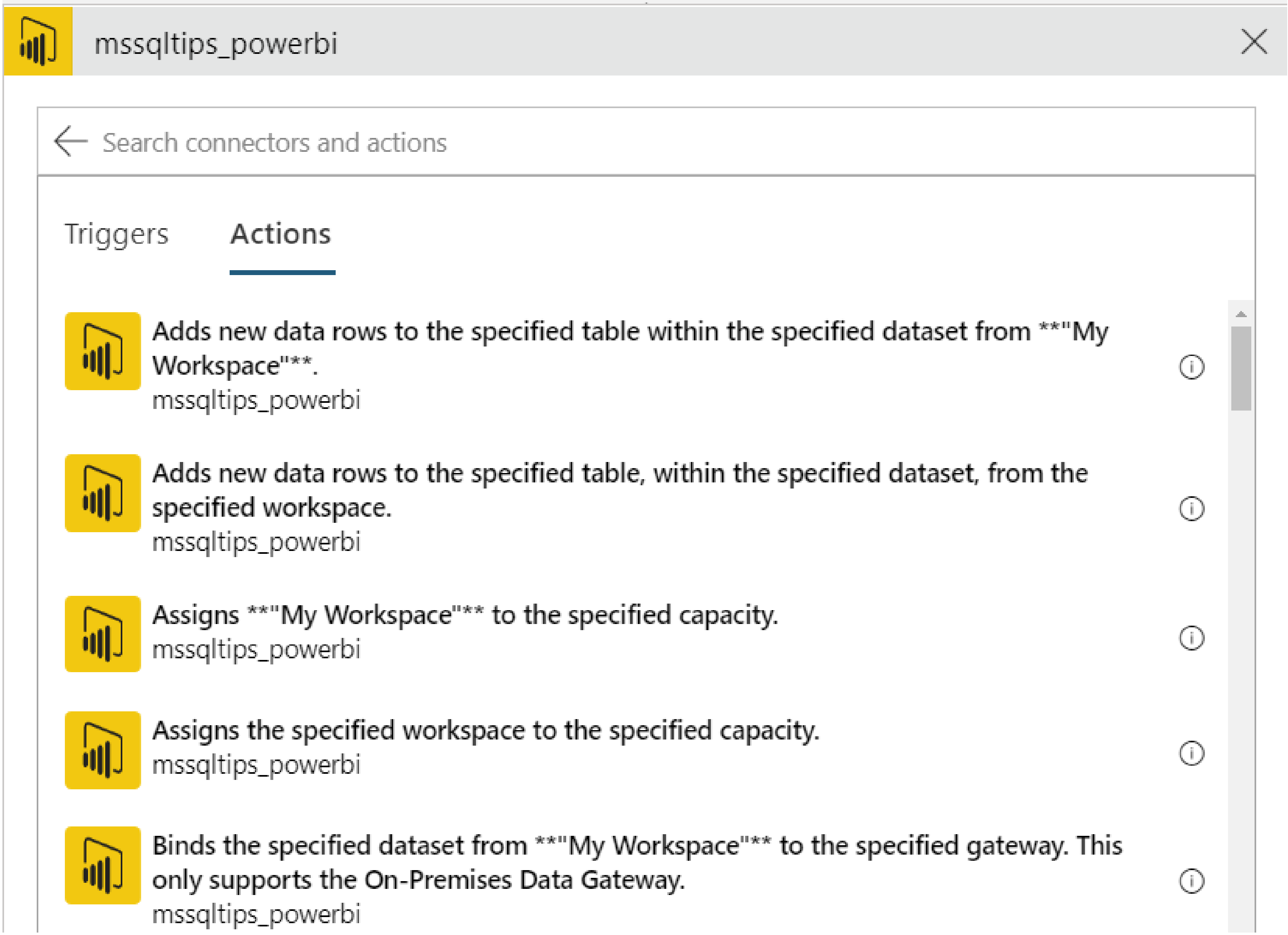
For each

azure logic apps

extensibility through customer connectors

e.g. an API and its swagger file

each API method can be translated into an action



mssqltips_powerbi

Search connectors and actions

Adds new data rows to the specified table within the specified dataset from **"My Workspace"**.
mssqltips_powerbi

Adds new data rows to the specified table, within the specified dataset, from the specified workspace.
mssqltips_powerbi

Assigns **"My Workspace"** to the specified capacity.
mssqltips_powerbi

Assigns the specified workspace to the specified capacity.
mssqltips_powerbi

Binds the specified dataset from **"My Workspace"** to the specified gateway. This only supports the On-Premises Data Gateway.
mssqltips_powerbi

ae architects for business & ict

DEMO TIME

architects
for business
& ict

azure data factory

ae

architects
for business
& ict

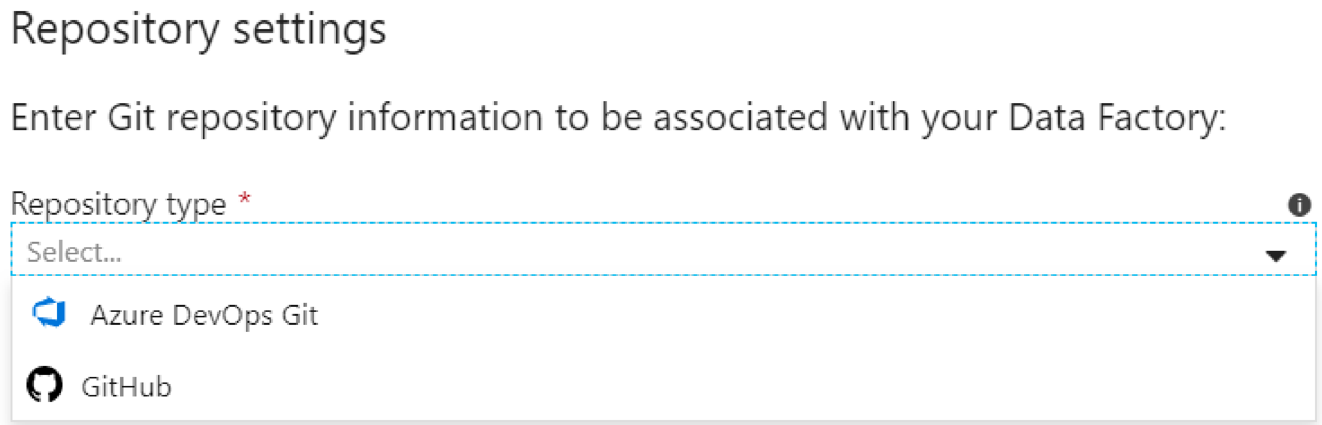
data factory v2

data integration service

ELT/ETL without writing code

scalable data pipelines

git integration



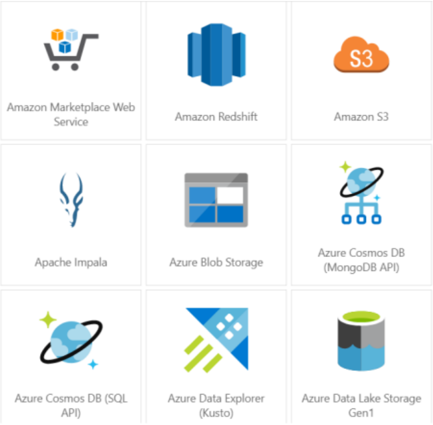
Repository settings

Enter Git repository information to be associated with your Data Factory:

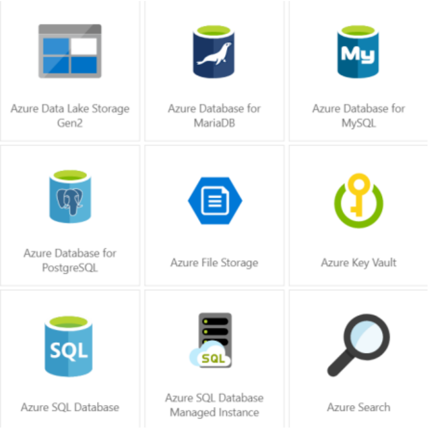
Repository type *

Select...

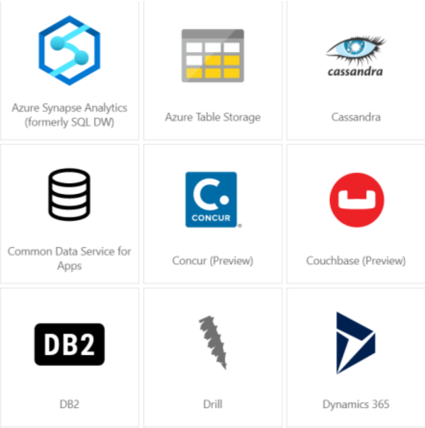
linked services



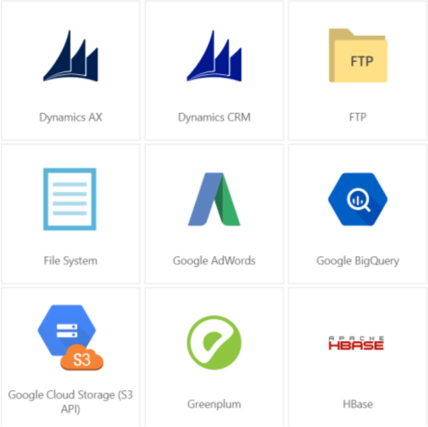
Amazon Marketplace Web Service	Amazon Redshift	Amazon S3
Apache Impala	Azure Blob Storage	Azure Cosmos DB (MongoDB API)
Azure Cosmos DB (SQL API)	Azure Data Explorer (Kusto)	Azure Data Lake Storage Gen1



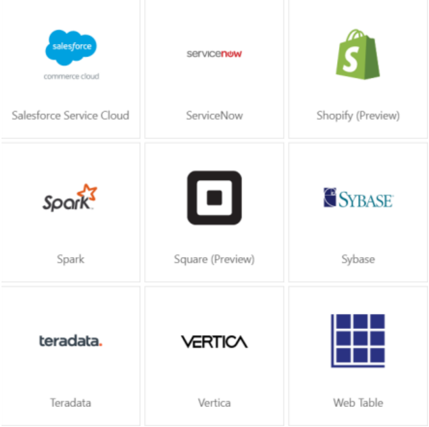
Azure Data Lake Storage Gen2	Azure Database for MariaDB	Azure Database for MySQL
Azure Database for PostgreSQL	Azure File Storage	Azure Key Vault
Azure SQL Database	Azure SQL Database Managed Instance	Azure Search



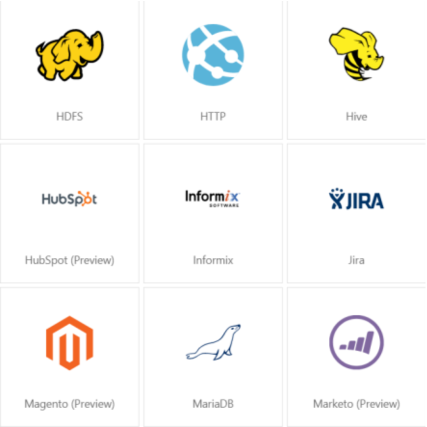
Azure Synapse Analytics (formerly SQL DW)	Azure Table Storage	Cassandra
Common Data Service for Apps	Concur (Preview)	Couchbase (Preview)
DB2	Drill	Dynamics 365



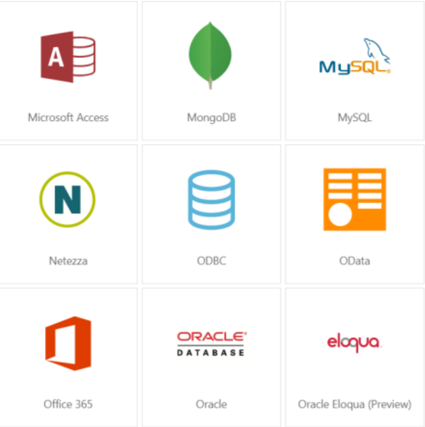
Dynamics AX	Dynamics CRM	FTP
File System	Google AdWords	Google BigQuery
Google Cloud Storage (S3 API)	Greenplum	HBase



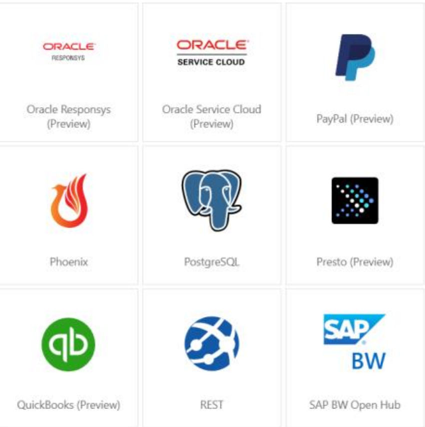
Salesforce Service Cloud	ServiceNow	Shopify (Preview)
Spark	Square (Preview)	Sybase
Teradata	Vertica	Web Table



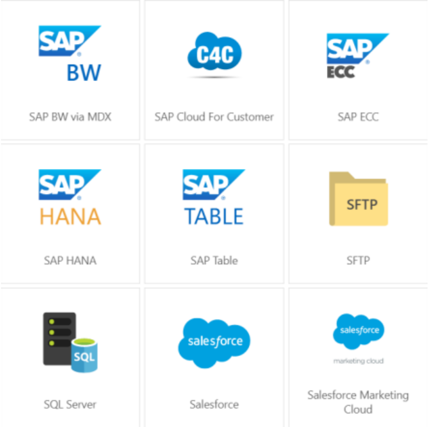
HDFS	HTTP	Hive
HubSpot (Preview)	Informix	Jira
Magento (Preview)	MariaDB	Marketo (Preview)



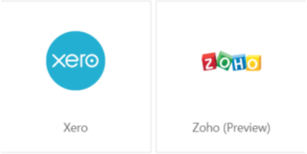
Microsoft Access	MongoDB	MySQL
Netezza	ODBC	OData
Office 365	Oracle	Oracle Eloqua (Preview)



Oracle Responsys (Preview)	Oracle Service Cloud (Preview)	PayPal (Preview)
Phoenix	PostgreSQL	Presto (Preview)
QuickBooks (Preview)	REST	SAP BW Open Hub



SAP BW via MDX	SAP Cloud For Customer	SAP ECC
SAP HANA	SAP Table	SFTP
SQL Server	Salesforce	Salesforce Marketing Cloud



Xero	Zoho (Preview)

architects
for business
& ict

datasets

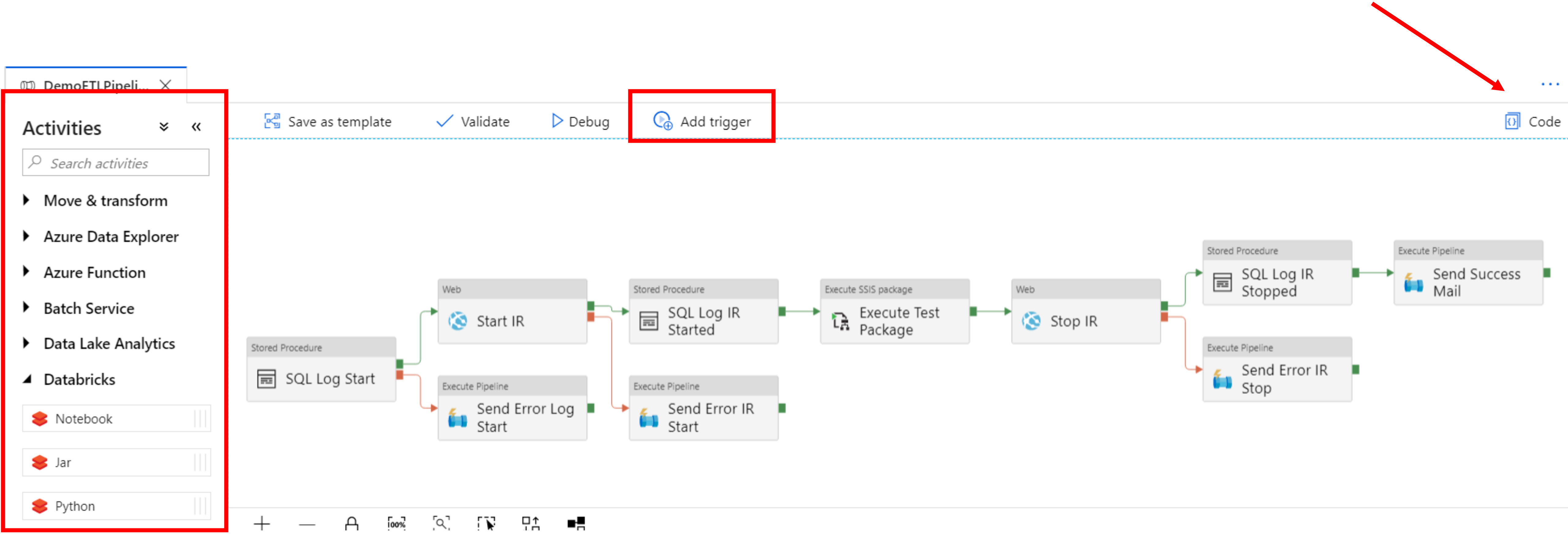
What type of data? Where is it located?
What is the schema?

SQL

dbo.dimEmployee

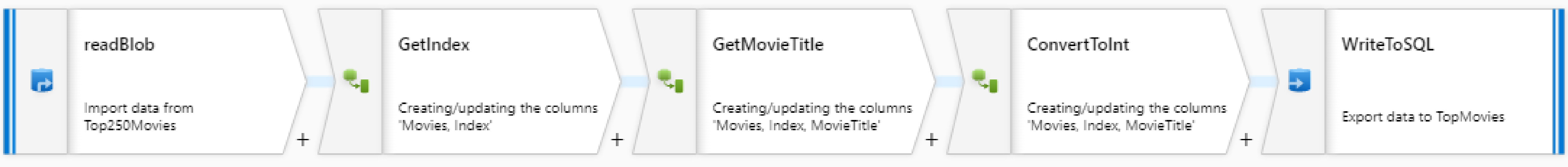
/clean/2019/12/10/logs.csv

pipelines



DemoETLPipeli...
Activities
Search activities
Move & transform
Azure Data Explorer
Azure Function
Batch Service
Data Lake Analytics
Databricks
Notebook
Jar
Python
Save as template
Validate
Debug
Add trigger
Code
Stored Procedure
SQL Log Start
Web
Start IR
Execute Pipeline
Send Error Log Start
Stored Procedure
SQL Log IR Started
Execute Pipeline
Send Error IR Start
Execute SSIS package
Execute Test Package
Web
Stop IR
Stored Procedure
SQL Log IR Stopped
Execute Pipeline
Send Error IR Stop
Execute Pipeline
Send Success Mail

mapping data flow



readBlob

Import data from Top250Movies

GetIndex

Creating/updating the columns 'Movies, Index'

GetMovieTitle

Creating/updating the columns 'Movies, Index, MovieTitle'

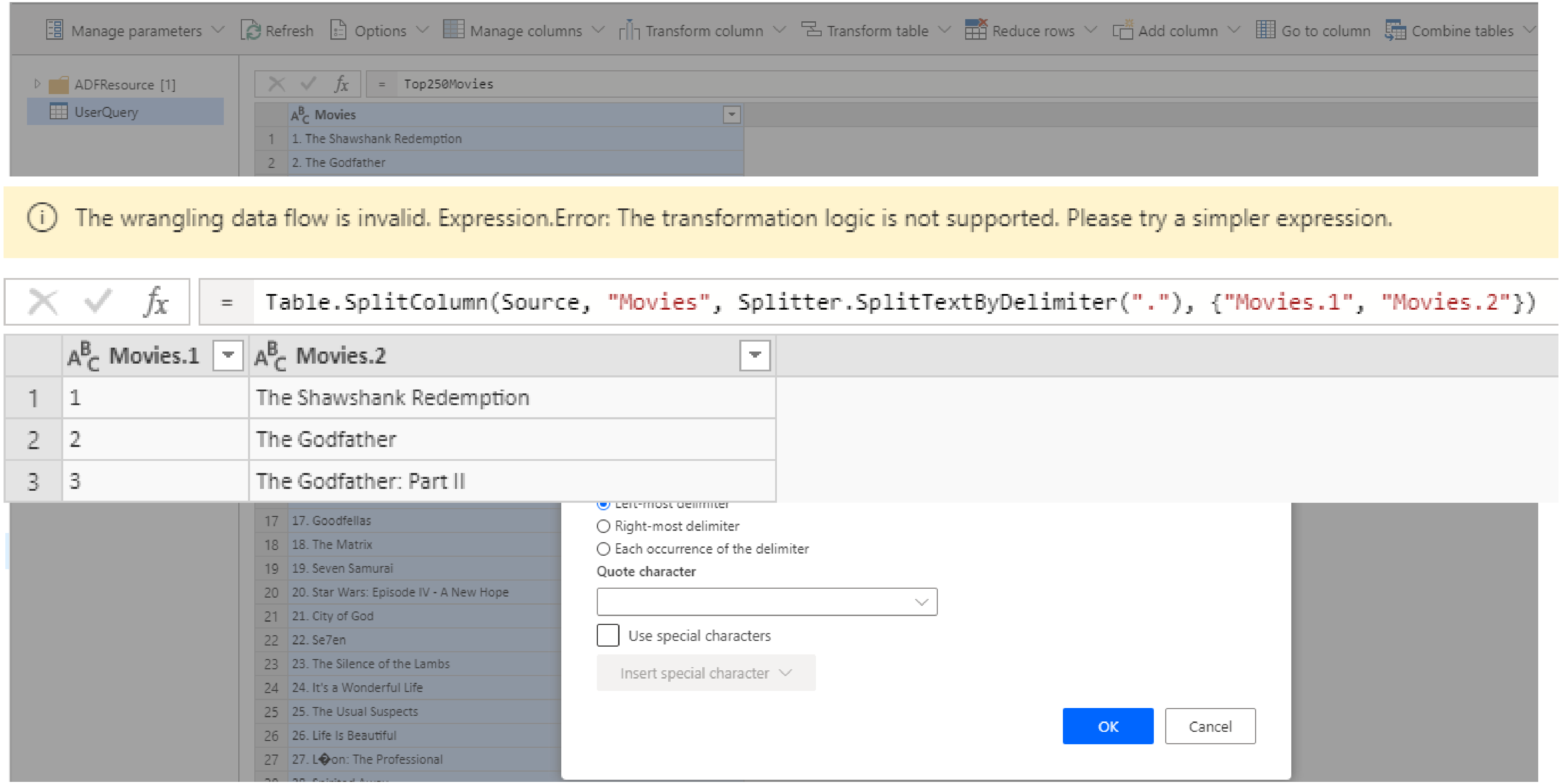
ConvertToInt

Creating/updating the columns 'Movies, Index, MovieTitle'

WriteToSQL

Export data to TopMovies

wrangling data flow



```
= Top250Movies
```

	Movies
1	1. The Shawshank Redemption
2	2. The Godfather

ⓘ The wrangling data flow is invalid. Expression.Error: The transformation logic is not supported. Please try a simpler expression.

```
= Table.SplitColumn(Source, "Movies", Splitter.SplitTextByDelimiter("."), {"Movies.1", "Movies.2"})
```

	Movies.1	Movies.2
1	1	The Shawshank Redemption
2	2	The Godfather
3	3	The Godfather: Part II

17	17. Goodfellas
18	18. The Matrix
19	19. Seven Samurai
20	20. Star Wars: Episode IV - A New Hope
21	21. City of God
22	22. Se7en
23	23. The Silence of the Lambs
24	24. It's a Wonderful Life
25	25. The Usual Suspects
26	26. Life Is Beautiful
27	27. L�on: The Professional
28	28. Spirited Away

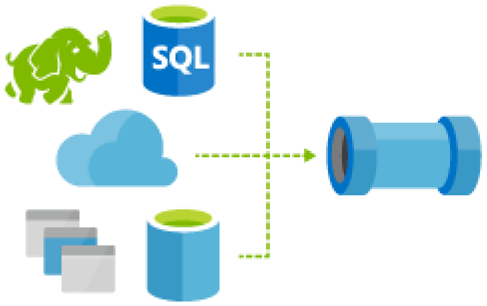
● Left-most delimiter
○ Right-most delimiter
○ Each occurrence of the delimiter

Quote character

□ Use special characters

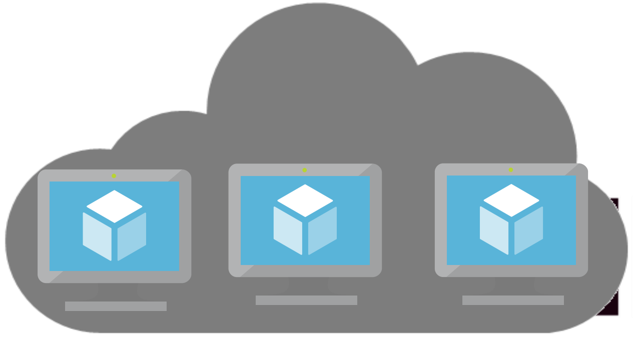
runtime

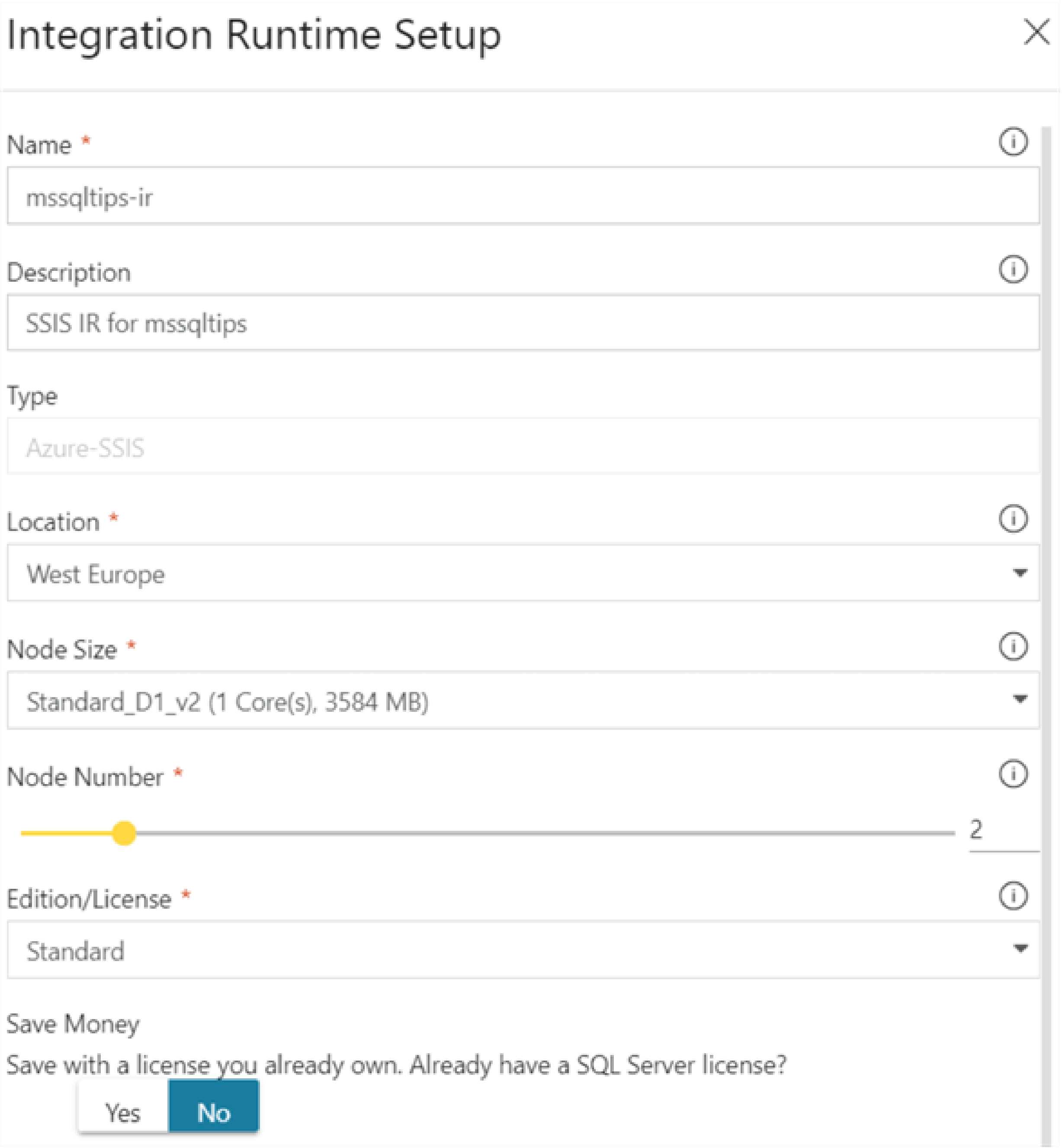
Azure Integration Runtime



Self-hosted Integration Runtime

Azure SSIS Integration Runtime





Integration Runtime Setup

Name *

mssqltips-ir

Description

SSIS IR for mssqltips

Type

Azure-SSIS

Location *

West Europe

Node Size *

Standard_D1_v2 (1 Core(s), 3584 MB)

Node Number *

2

Edition/License *

Standard

Save Money

Save with a license you already own. Already have a SQL Server license?

Yes [✓ No]

DEMO TIME

architects
for business
& ict

CONCLUSION

ae®

architects
for business
& ict

conclusion

Azure Functions

- if you like coding
- event-based
- more suitable for smaller, specific tasks

Azure Databricks

- big data solution
- for data engineering / data science

ae® architects for business & ict

conclusion

Logic Apps

- easy to build workflows
- lots of connectors
- event-based

Azure Data Factory

- most comprehensive and flexible
- visual tool
- big & small data