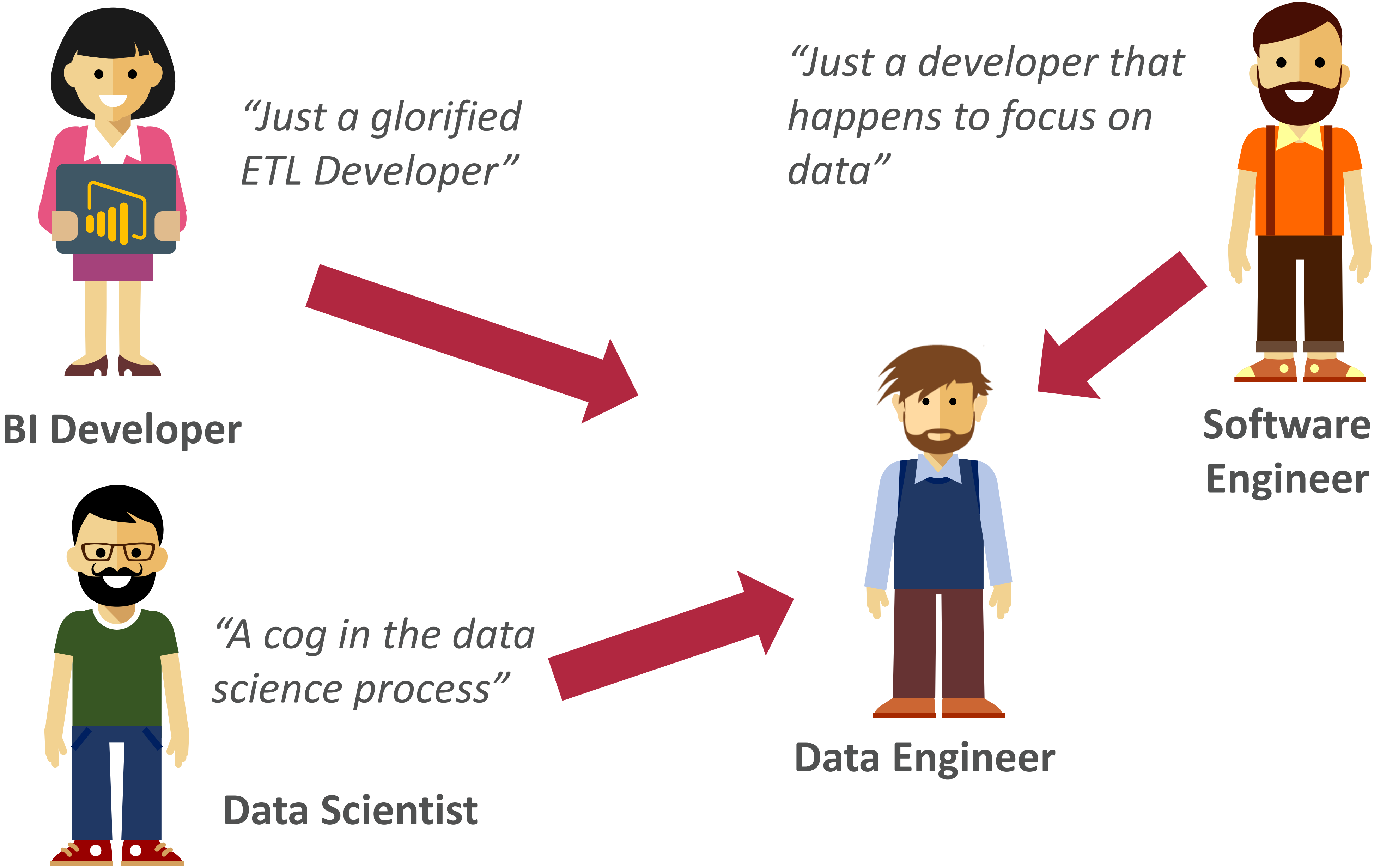
www.advancinganalytics.co.uk

A DATA ENGINEERS TOOLKIT

@ADVANCINGANALYTICS @ADVANALYTICSUK /ADVANCING ANALYTICS

SO WHAT IS DATA ENGINEERING?



BI Developer

"Just a glorified ETL Developer"

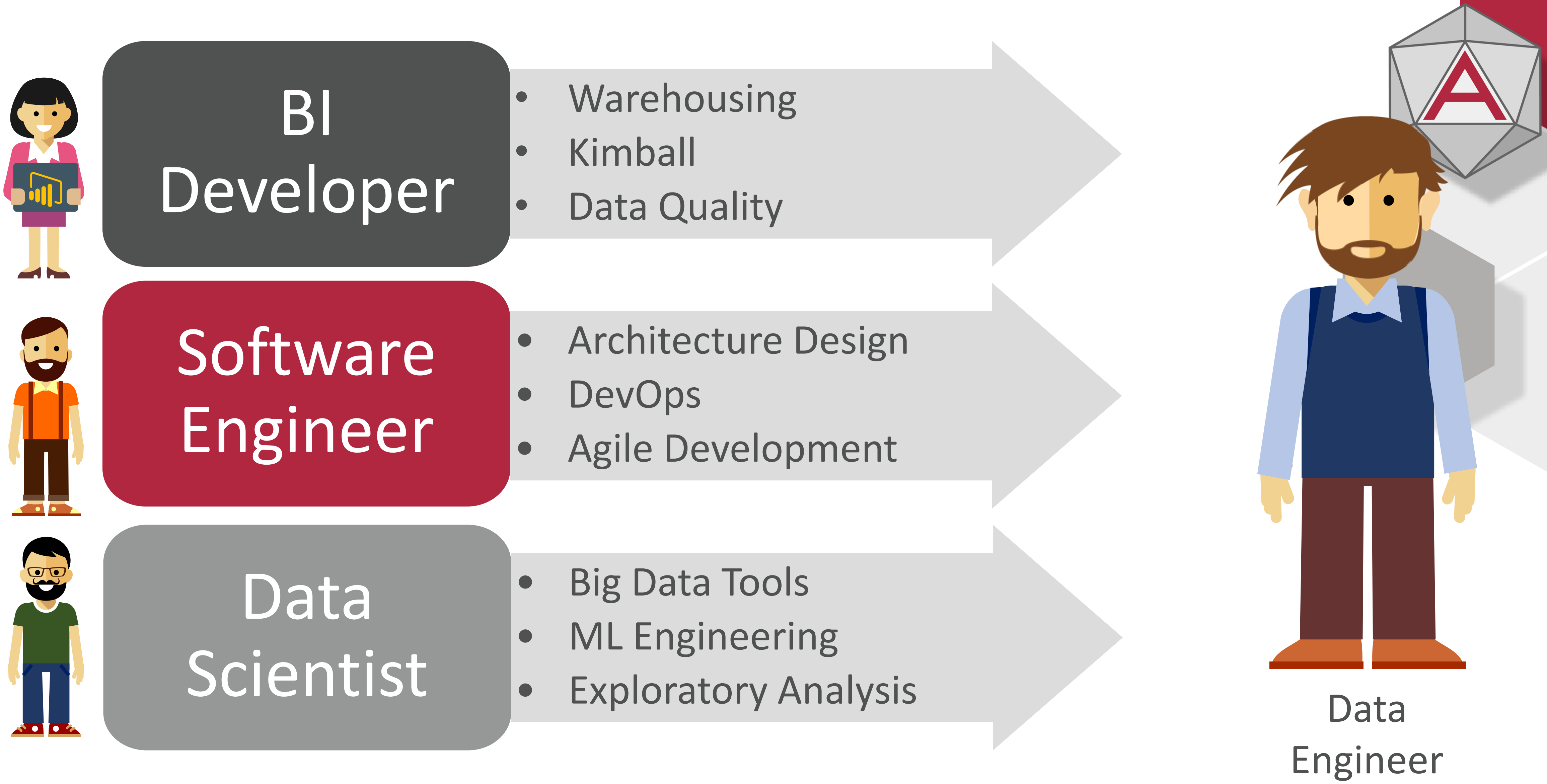
"Just a developer that happens to focus on data"

Software Engineer

Data Engineer

"A cog in the data science process"

Data Scientist



BI Developer

- Warehousing
- Kimball
- Data Quality

Software Engineer

- Architecture Design
- DevOps
- Agile Development

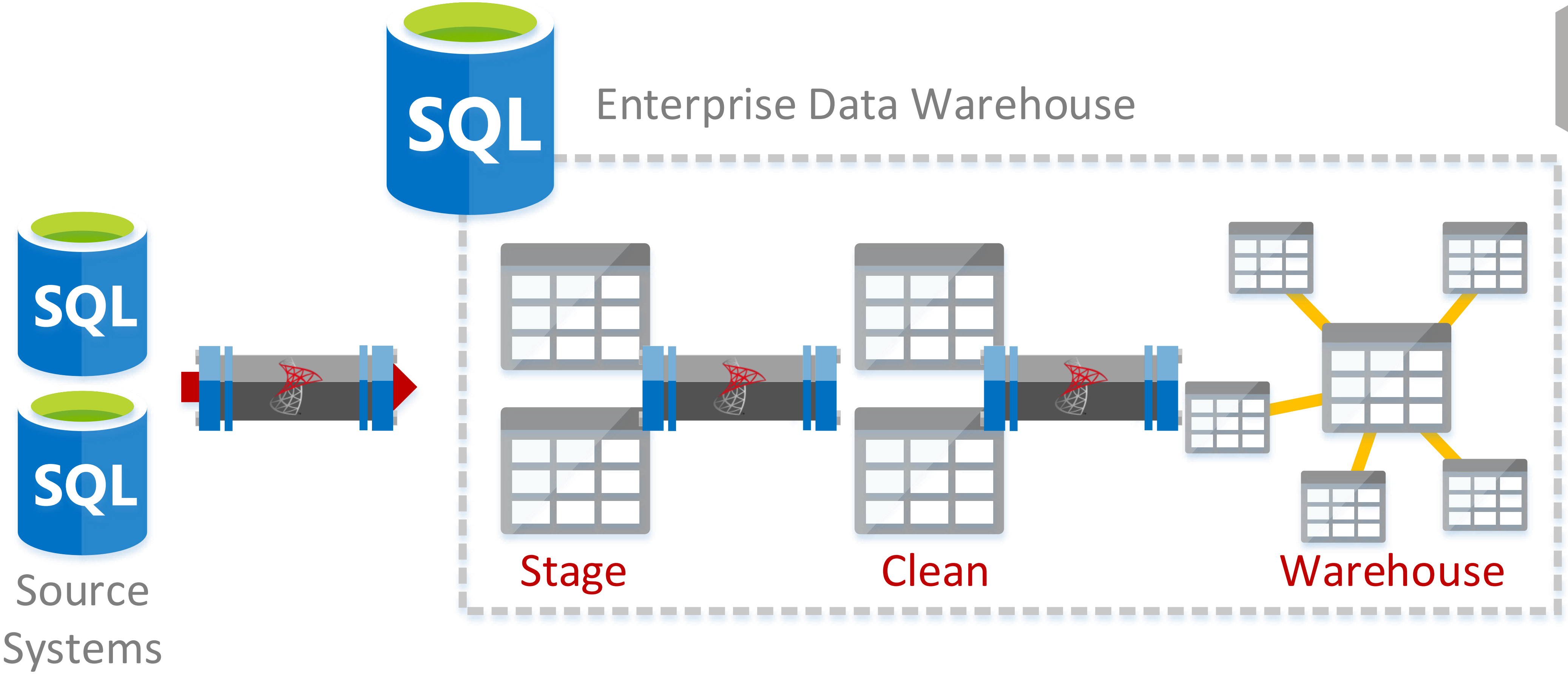
Data Scientist

- Big Data Tools
- ML Engineering
- Exploratory Analysis

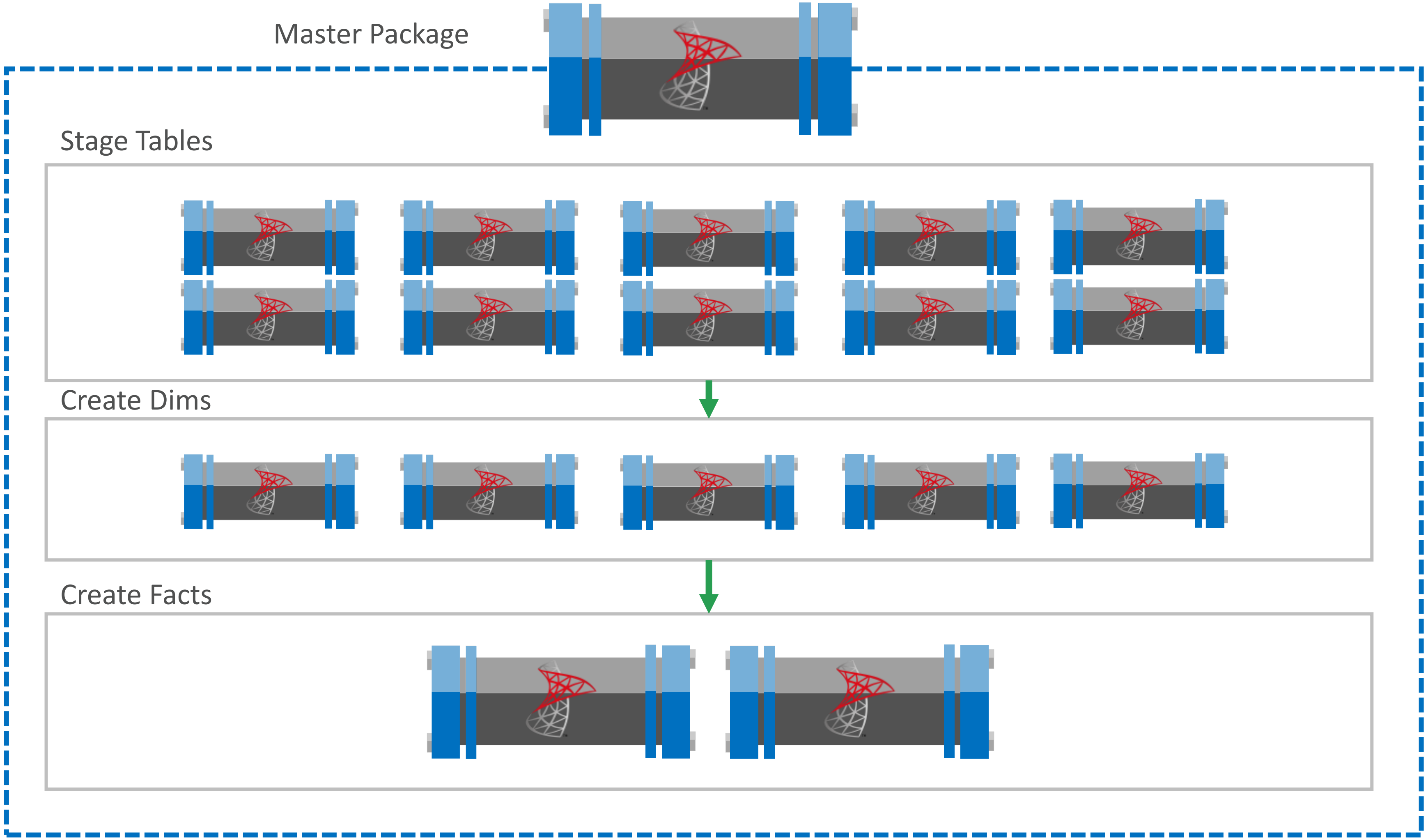
Data Engineer

EVOLVING ARCHITECTURES

THE OLD WORLD



PACKAGE ORCHESTRATION

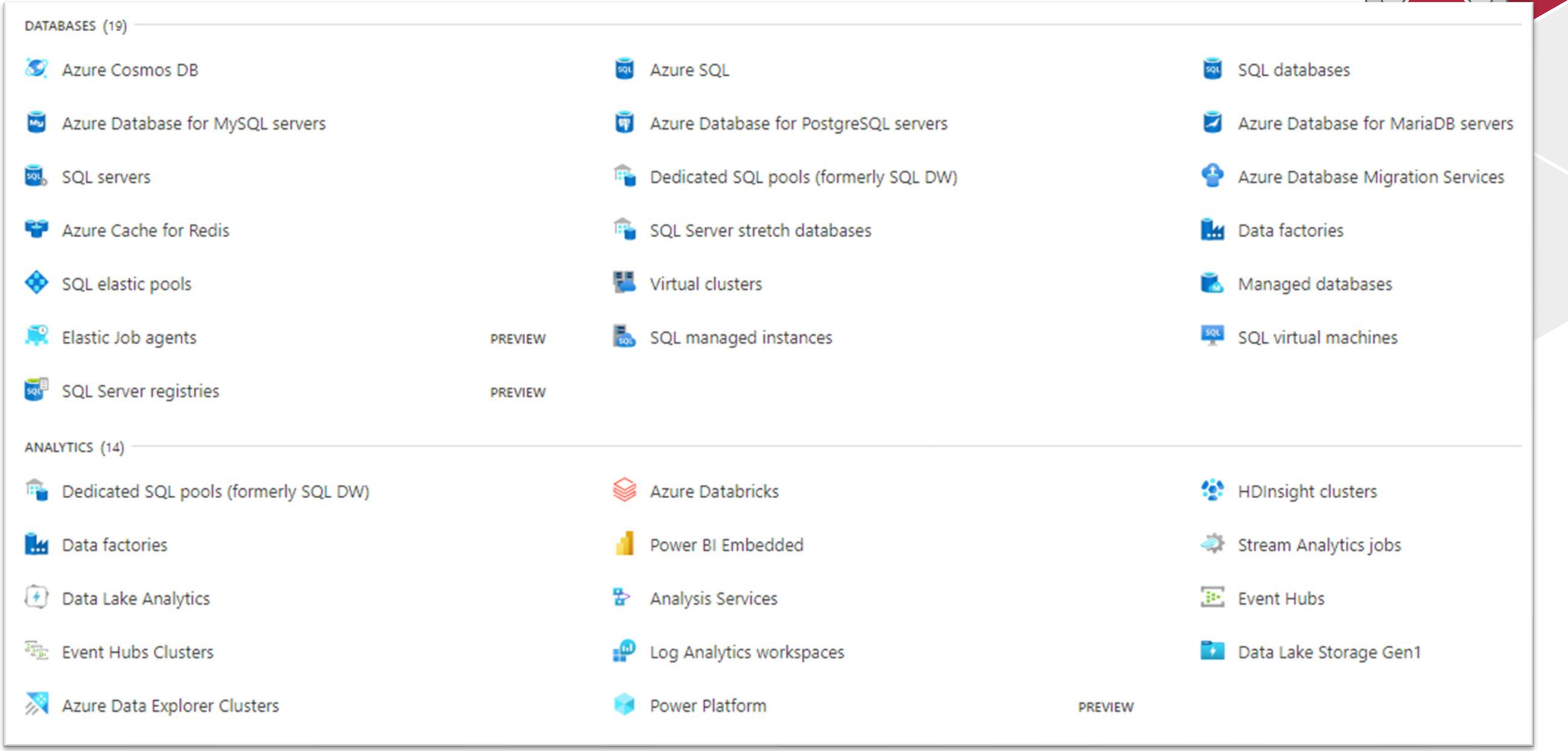


OUR ETL TOOL CHECKLIST

- Low Dev Effort
- Many Use Cases
- Flexible Formats
- Elastic Scaling
- Agile
- Support

FAILED

SO MOVE TO AZURE, IT'S EASY RIGHT?



DATABASES (19)

- Azure Cosmos DB
- Azure Database for MySQL servers
- SQL servers
- Azure Cache for Redis
- SQL elastic pools
- Elastic Job agents PREVIEW
- SQL Server registries PREVIEW
- Azure SQL
- Azure Database for PostgreSQL servers
- Dedicated SQL pools (formerly SQL DW)
- SQL Server stretch databases
- Virtual clusters
- SQL managed instances
- SQL databases
- Azure Database for MariaDB servers
- Azure Database Migration Services
- Data factories
- Managed databases
- SQL virtual machines

ANALYTICS (14)

- Dedicated SQL pools (formerly SQL DW)
- Data factories
- Data Lake Analytics
- Event Hubs Clusters
- Azure Data Explorer Clusters
- Azure Databricks
- Power BI Embedded
- Analysis Services
- Log Analytics workspaces
- Power Platform PREVIEW
- HDInsight clusters
- Stream Analytics jobs
- Event Hubs
- Data Lake Storage Gen1

MORE REALISTICALLY, WE HAVE SEVERAL KEY OPTIONS

Azure Data Factory

Visual cloud-native Orchestration. Excellent at moving data between sources

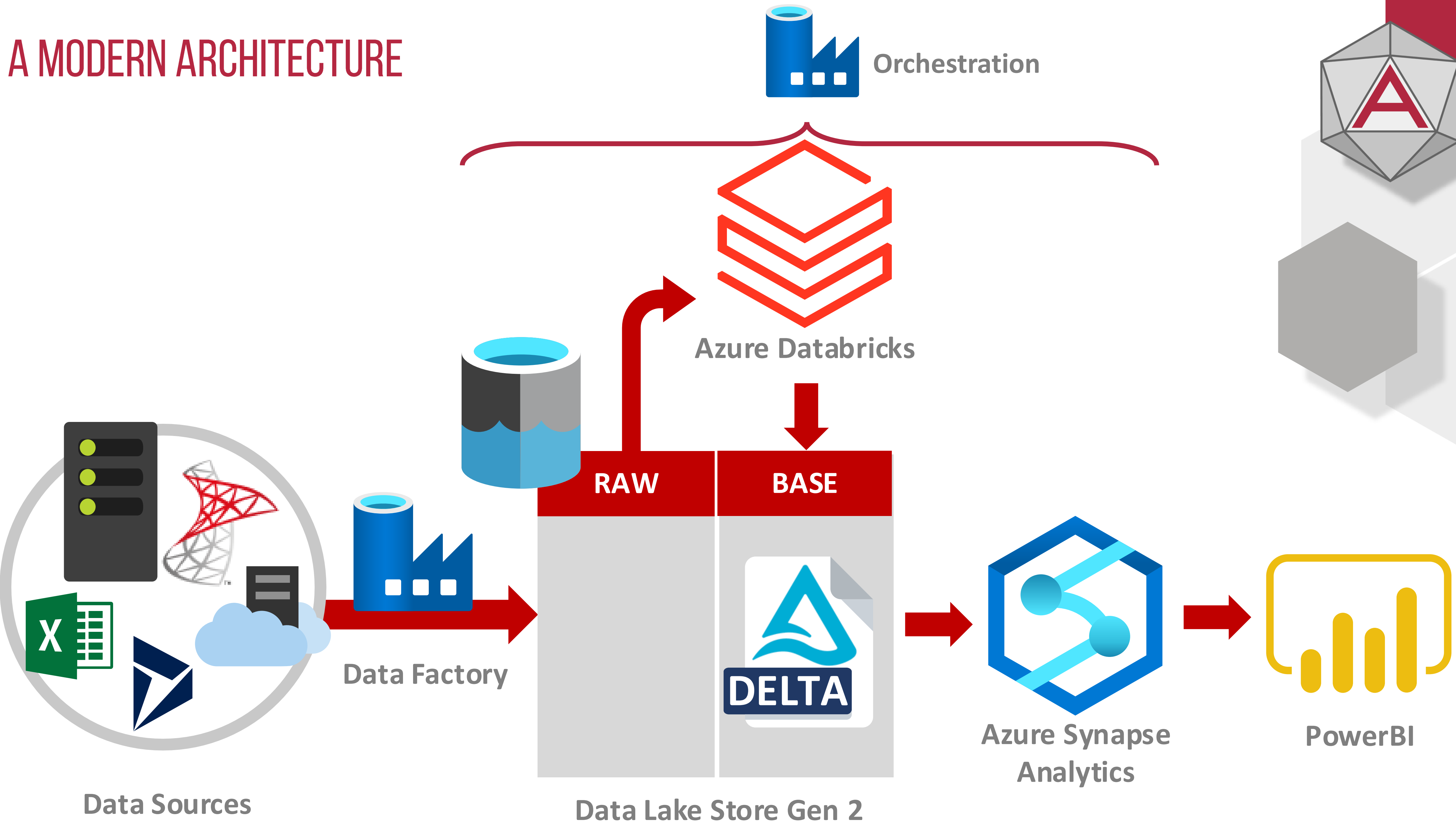
Databricks

Spark as a service, powerful multi-language big data engine. Supports SQL but best with Python/Scala

Azure Synapse Analytics

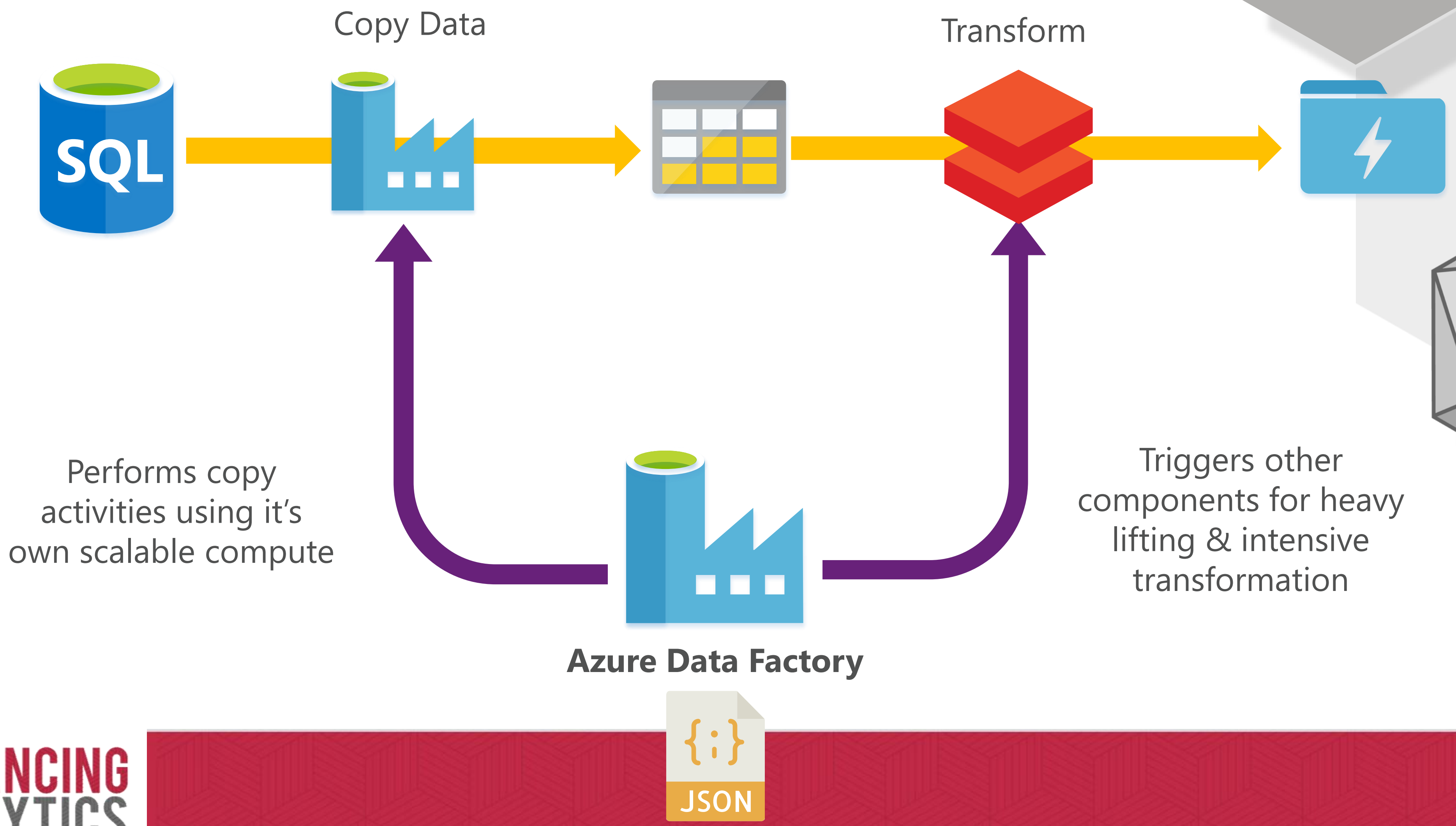
Suite of tools ranging from orchestration to engineering & full data warehousing

A MODERN ARCHITECTURE



ORCHESTRATION & INTEGRATION

AN ADF PRIMER



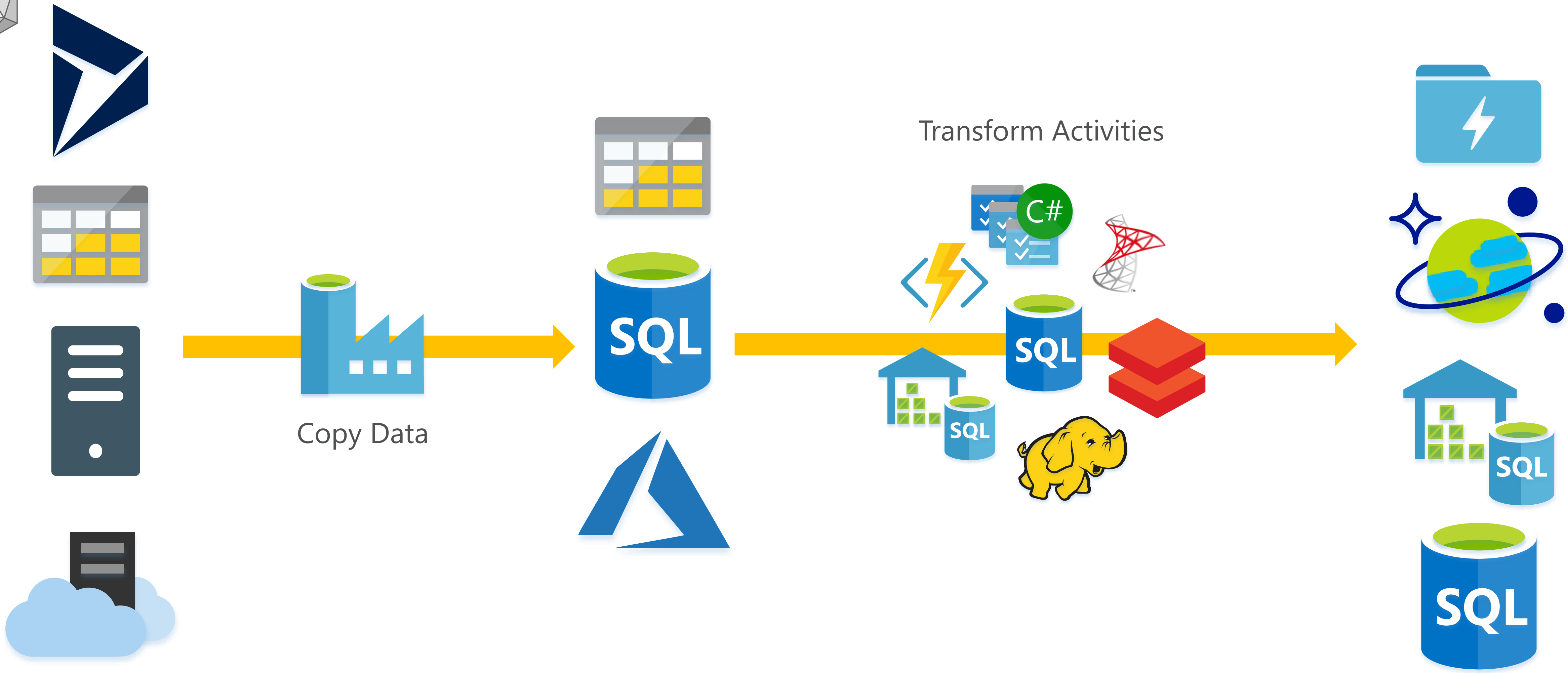
Copy Data

Transform

Performs copy activities using it's own scalable compute

Triggers other components for heavy lifting & intensive transformation

Azure Data Factory



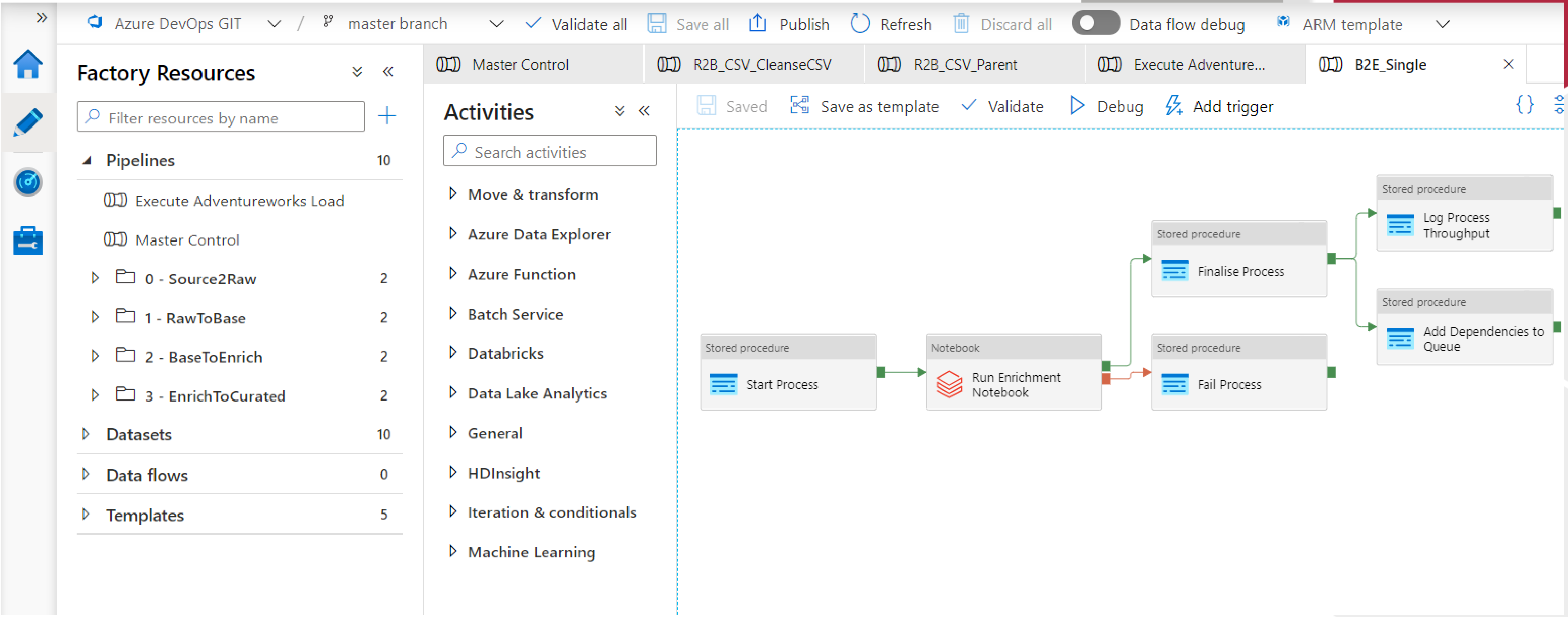
Transform Activities

Copy Data

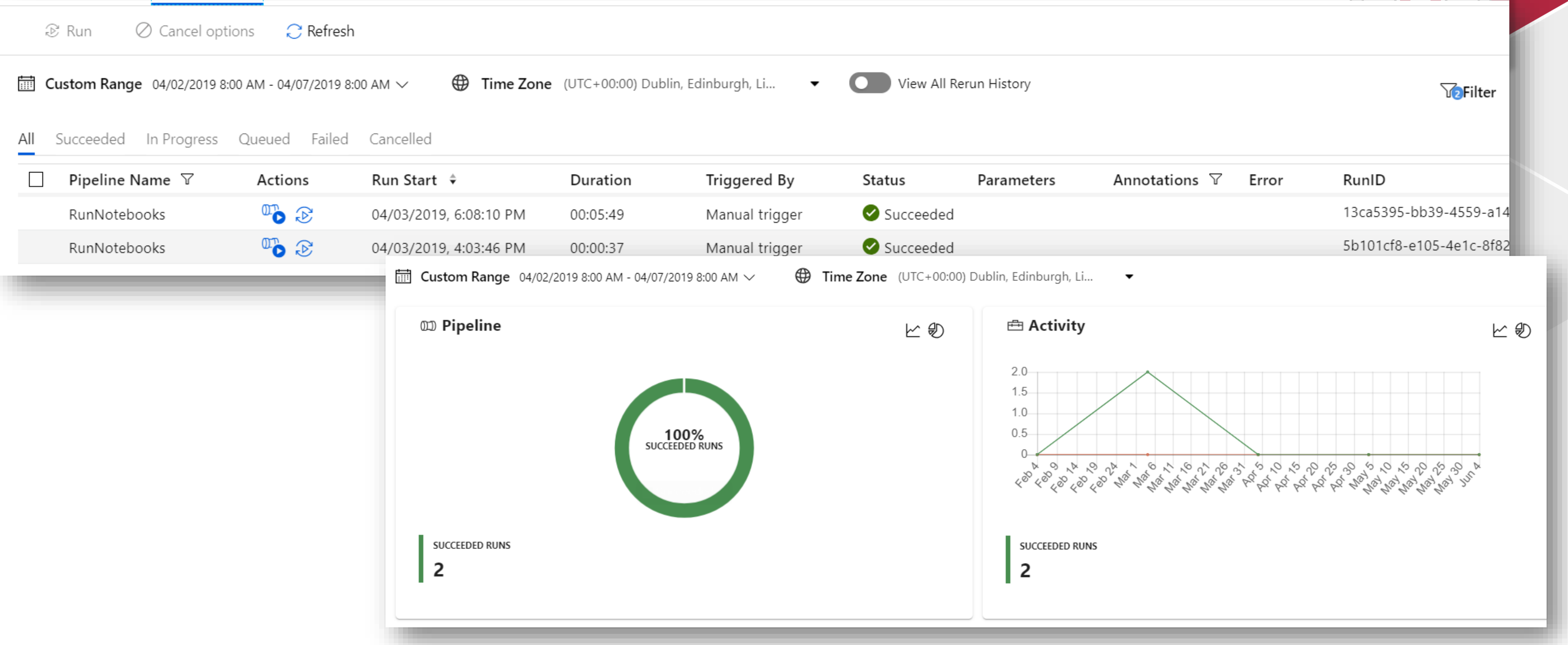
Sources

Destinations

DEVELOPER TOOLS



MONITORING



Custom Range 04/02/2019 8:00 AM - 04/07/2019 8:00 AM

Time Zone (UTC+00:00) Dublin, Edinburgh, Li...

View All Rerun History

□	Pipeline Name	Actions	Run Start	Duration	Triggered By	Status	Parameters	Annotations	Error	RunID
	RunNotebooks		04/03/2019, 6:08:10 PM	00:05:49	Manual trigger	✓ Succeeded				13ca5395-bb39-4559-a14
	RunNotebooks		04/03/2019, 4:03:46 PM	00:00:37	Manual trigger	✓ Succeeded				5b101cf8-e105-4e1c-8f82

Custom Range 04/02/2019 8:00 AM - 04/07/2019 8:00 AM

Time Zone (UTC+00:00) Dublin, Edinburgh, Li...

Pipeline

SUCCEEDED RUNS
2

Activity

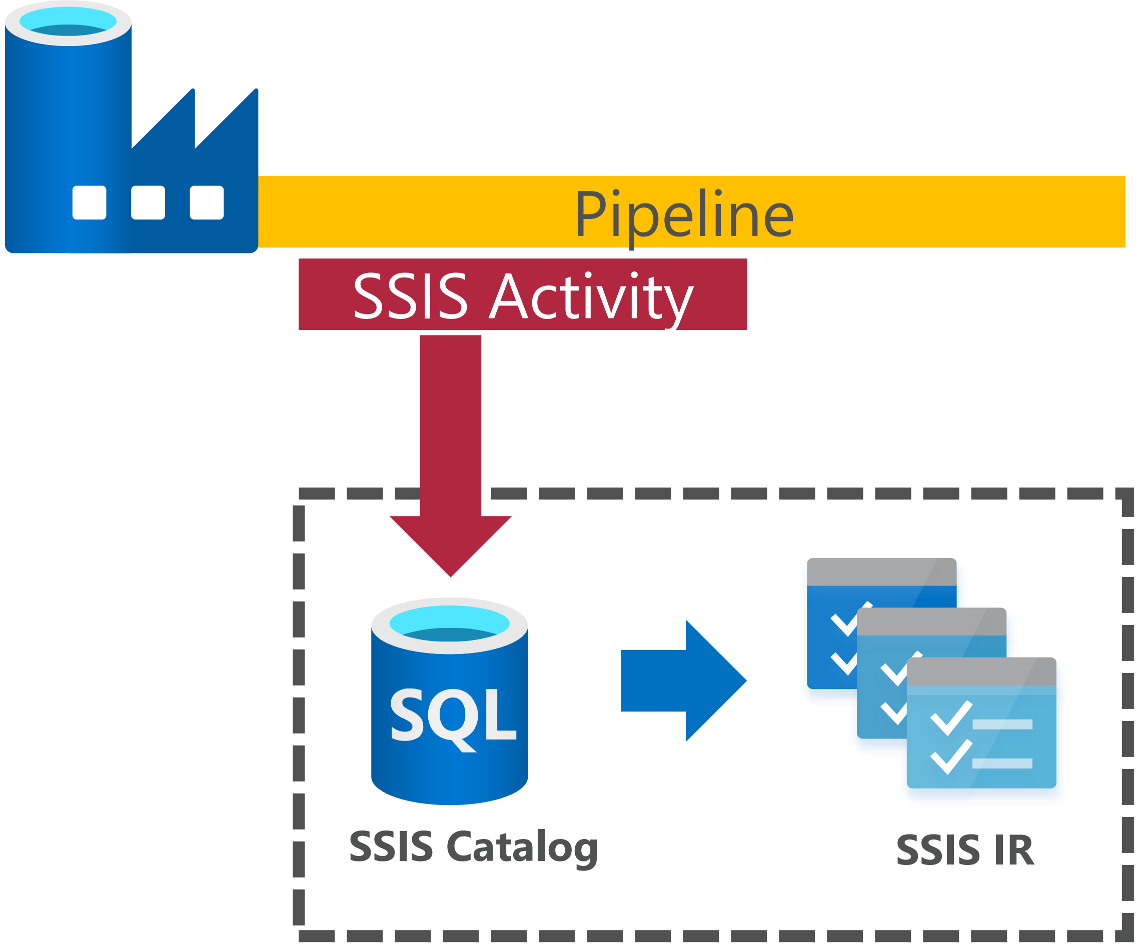
SUCCEEDED RUNS
2

ADVANCING ANALYTICS

A QUICK LOOK AT ADF

- The ADF Studio
- Object Concepts

SSIS – LIFT & SHIFT POTENTIAL



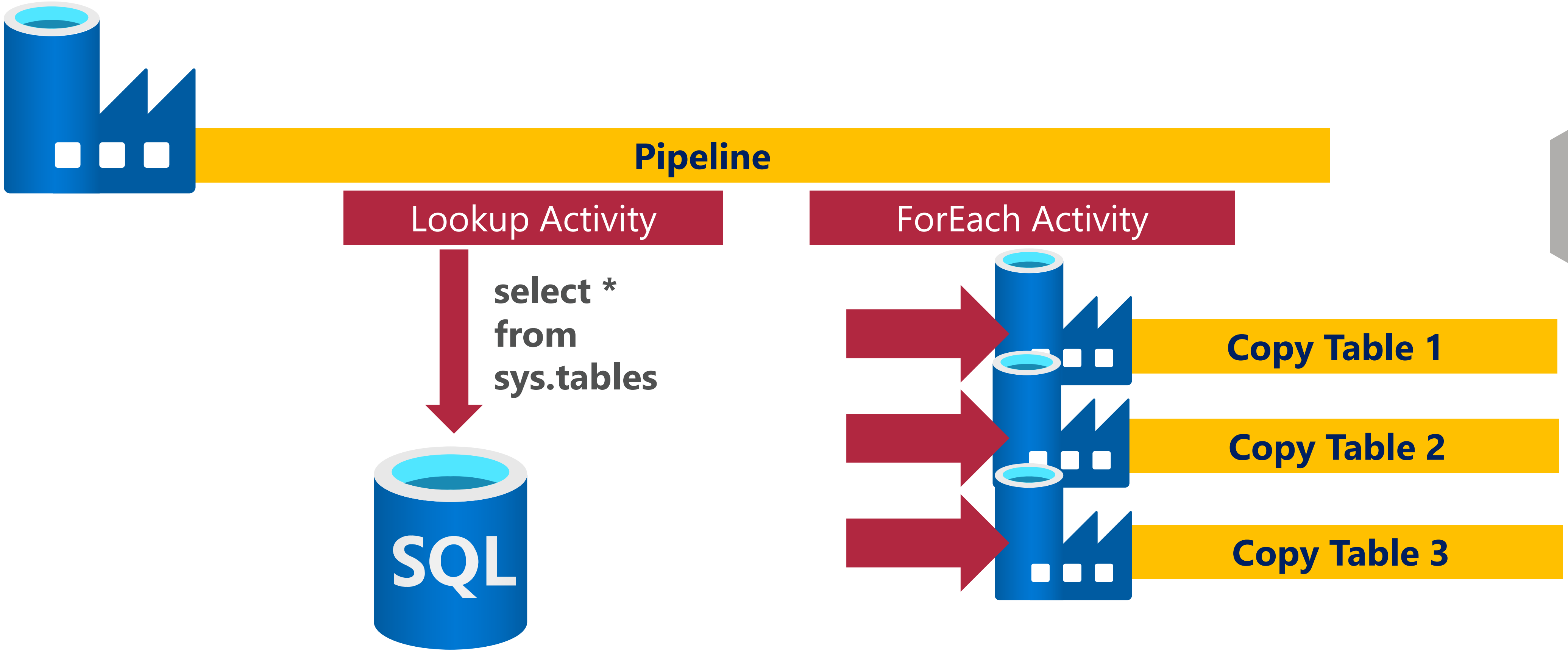
Pipeline

SSIS Activity

SSIS Catalog

SSIS IR

USE THE FLEXIBILITY OF DATA FACTORY



Pipeline

Lookup Activity

ForEach Activity

select *
from
sys.tables

Copy Table 1

Copy Table 2

Copy Table 3

DATA FACTORY RECAP

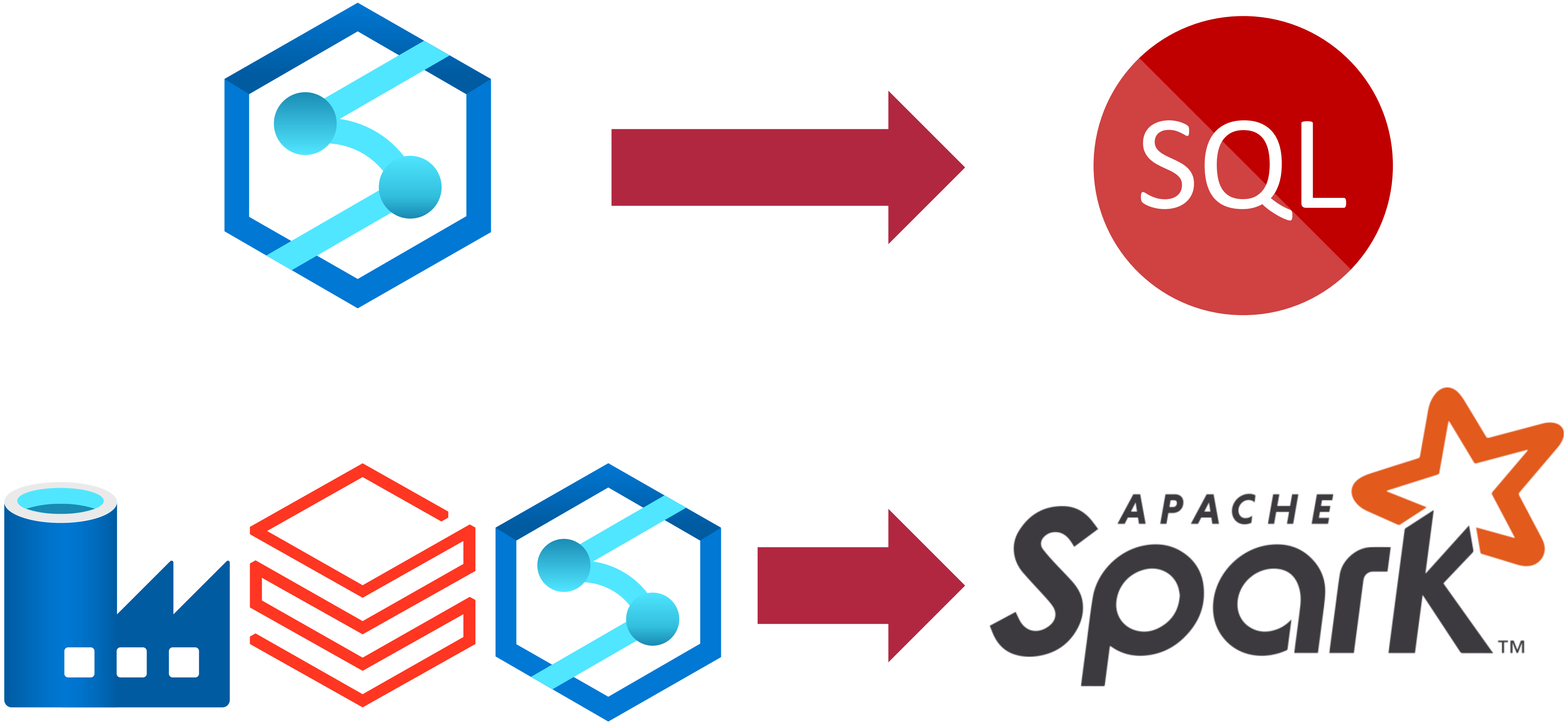
- Orchestrates all data workflows in Azure
- Best method of onboarding data to Azure
- Use parameters, forEach and child executions

- Low Dev Effort
- Many Use Cases
- Flexible Formats
- Elastic Scaling
- Agile
- Supportable

PASSED

DATA PROCESSING

COMPUTE APPROACHES IN AZURE

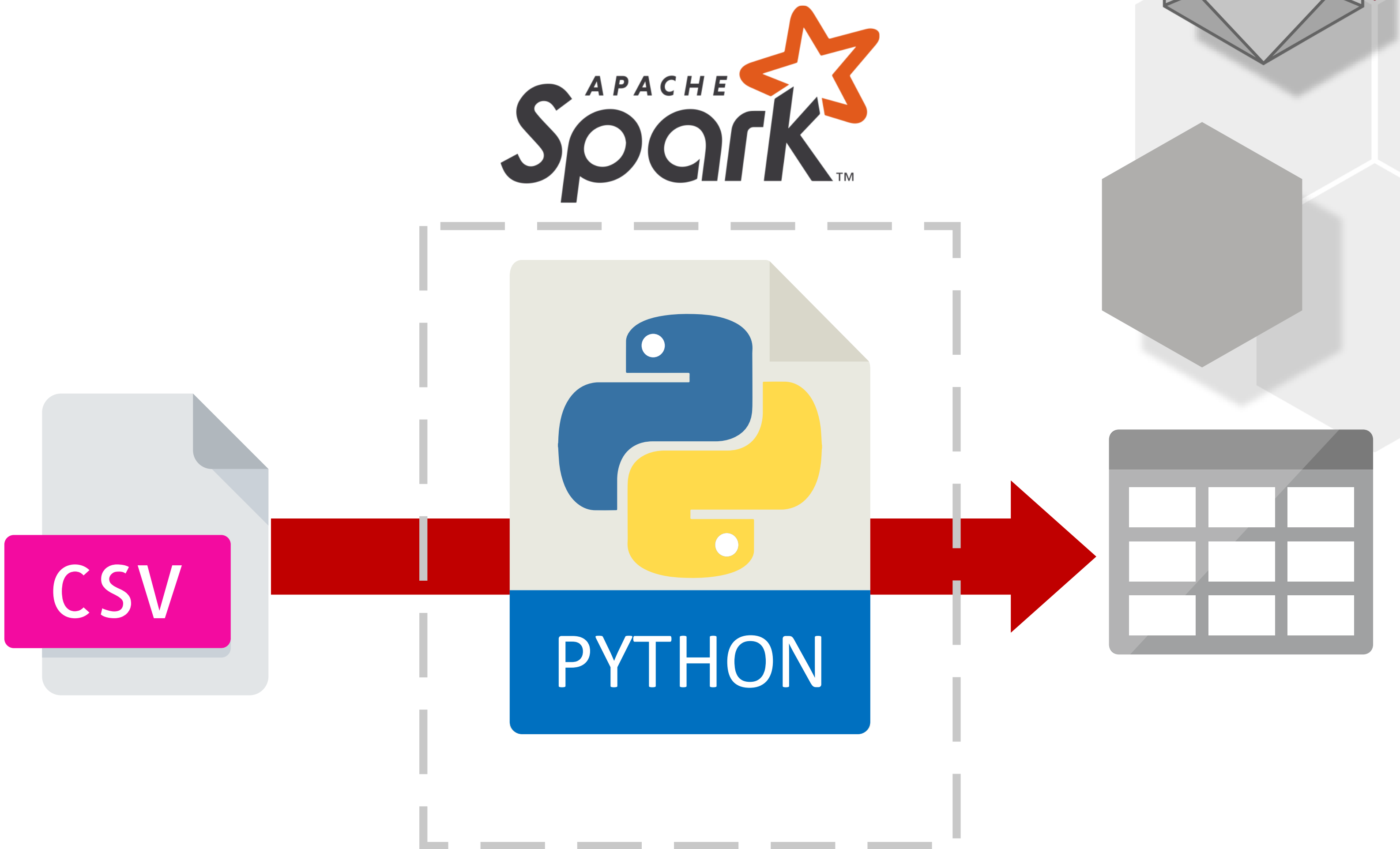


QUICK SPARK OVERVIEW

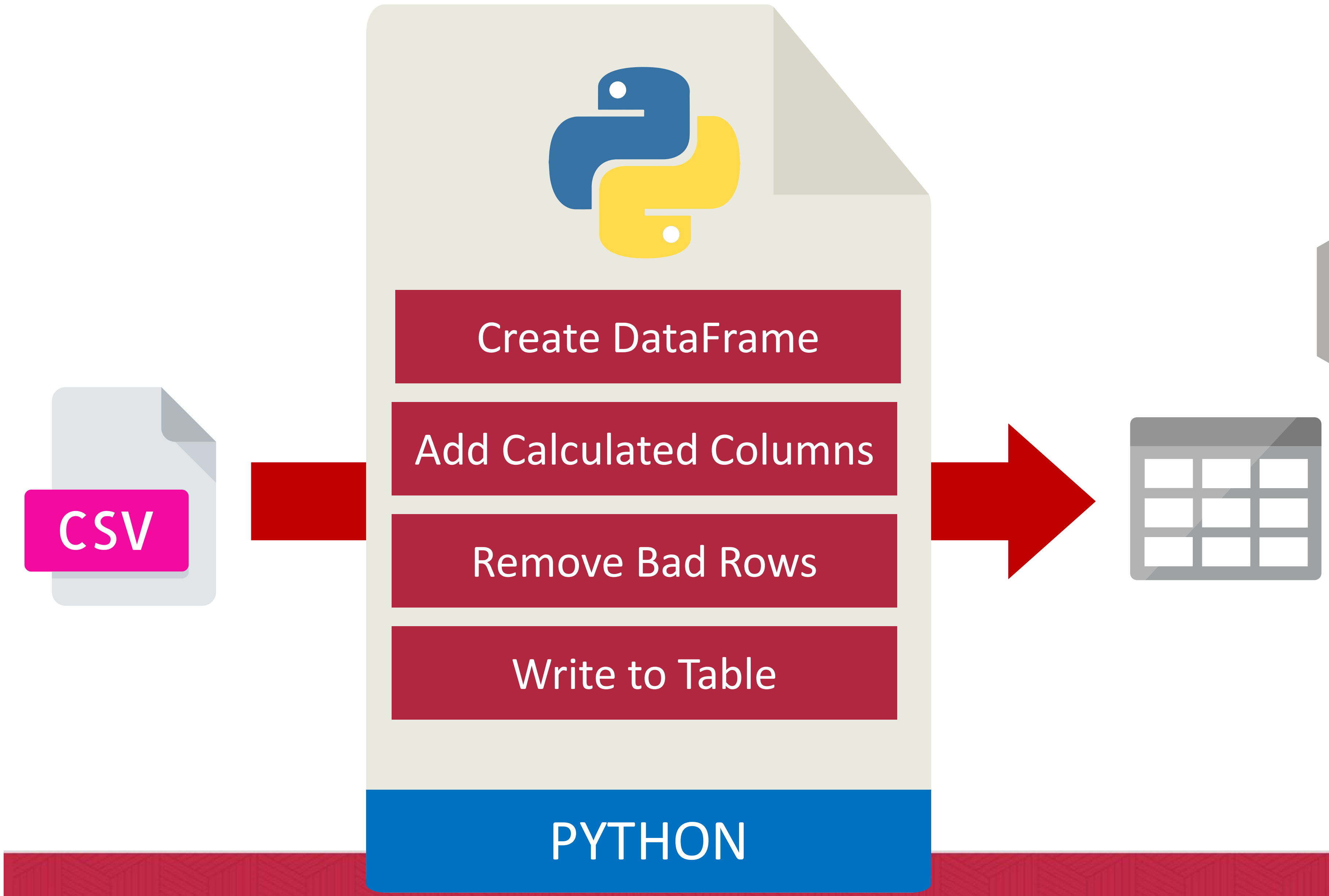
Spark is a distributed, scalable data processing engine.

It can query **structured** and **non-structured** data

You can use **Python**, **Scala**, **R**, **C#** or **SQL** to interact with it



SO WHAT?



CSV

Create DataFrame

Add Calculated Columns

Remove Bad Rows

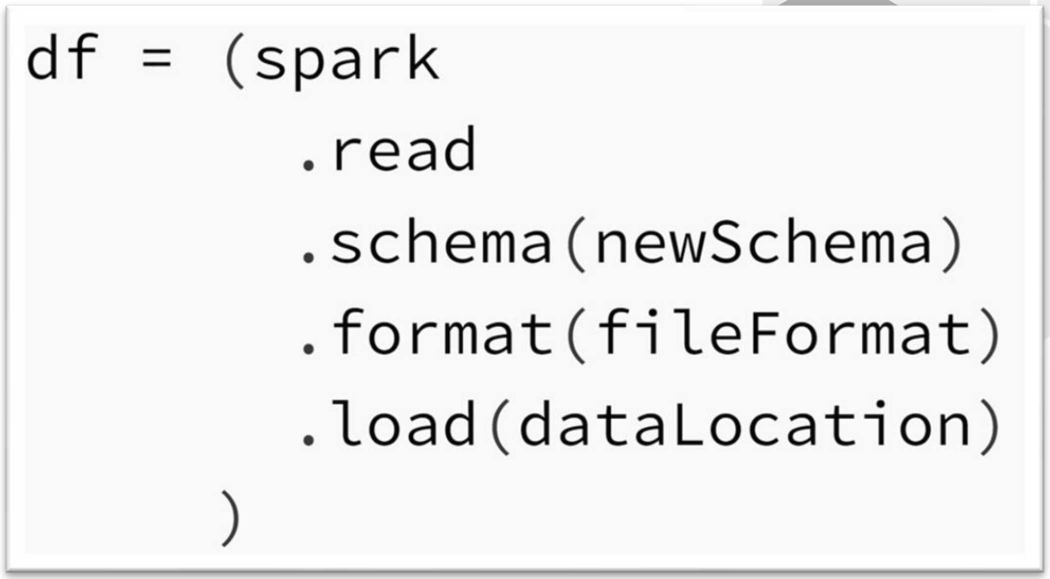
Write to Table

PYTHON

INSIDE A DATAFRAME

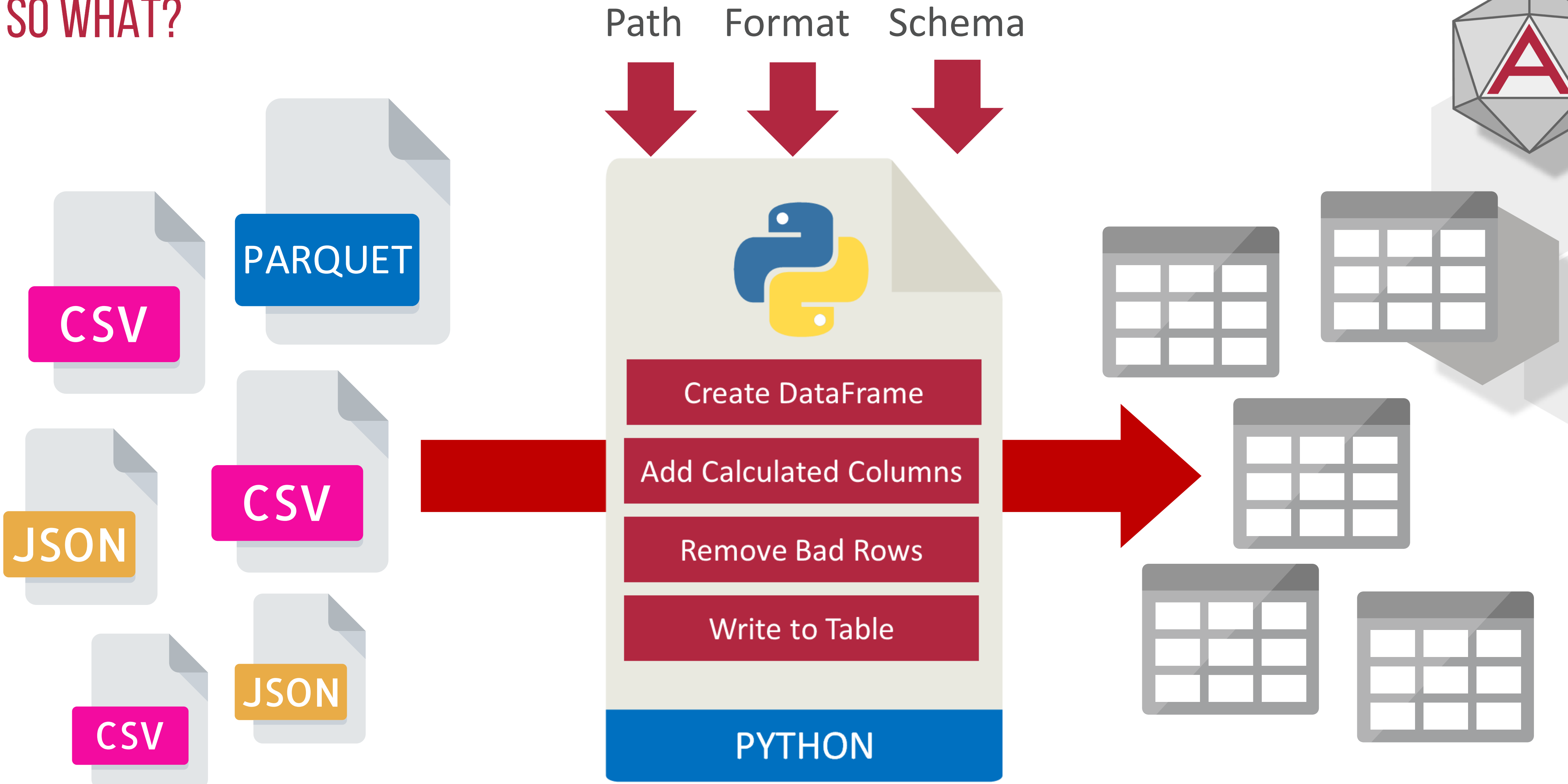
DataFrame

- Schema
- Format
- Location



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

SO WHAT?



Path

Format

Schema

Create DataFrame

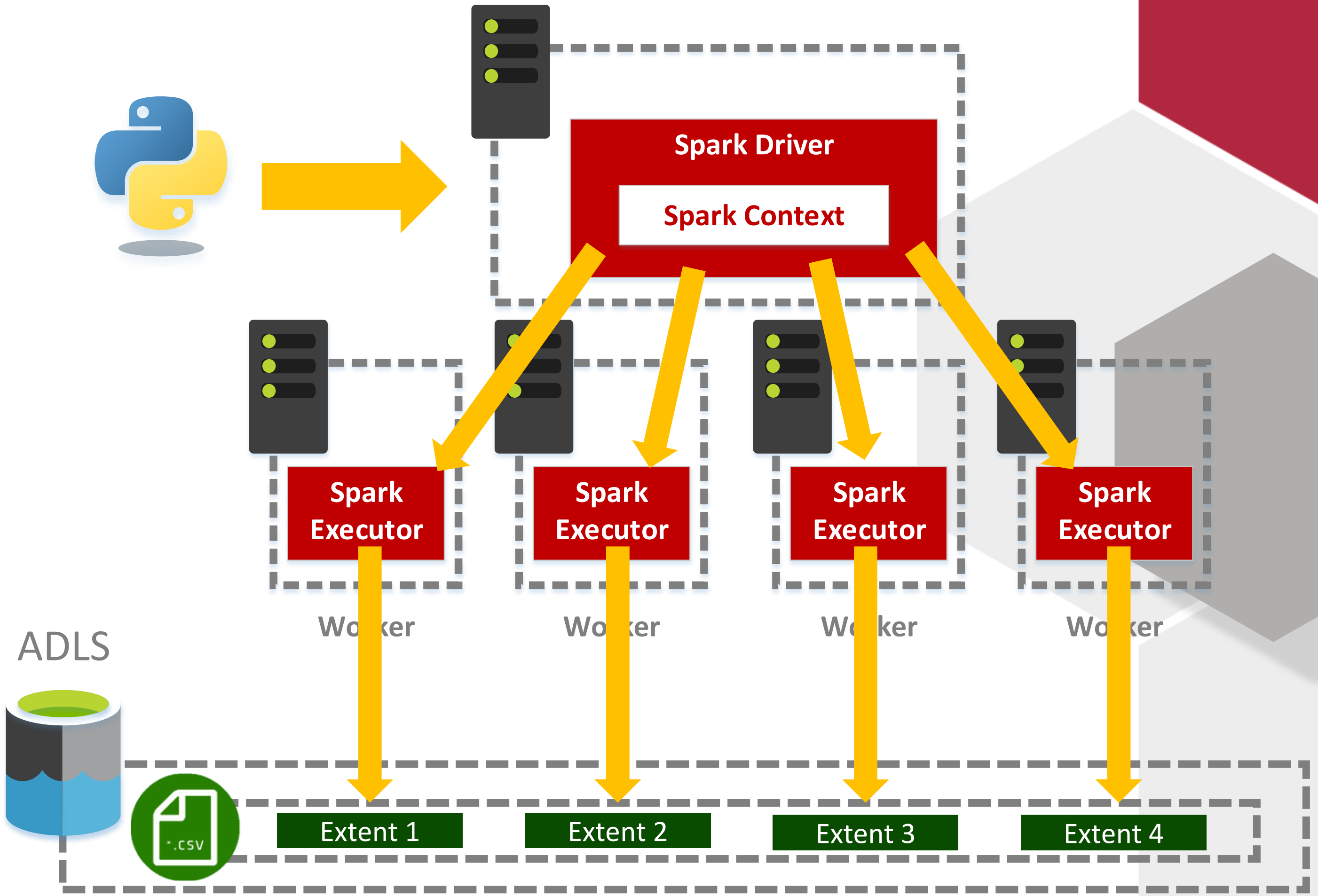
Add Calculated Columns

Remove Bad Rows

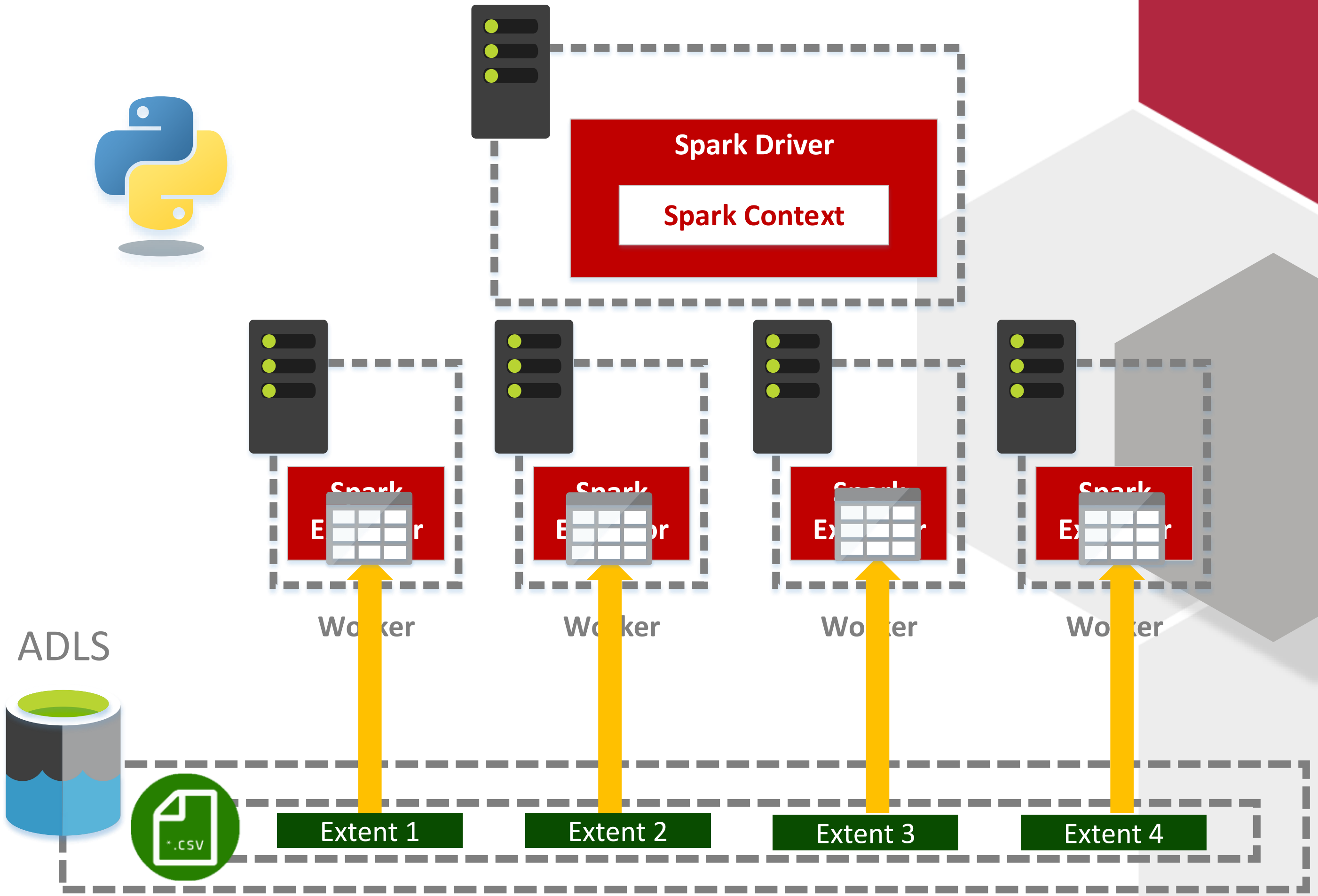
Write to Table

PYTHON

DISTRIBUTED COMPUTE

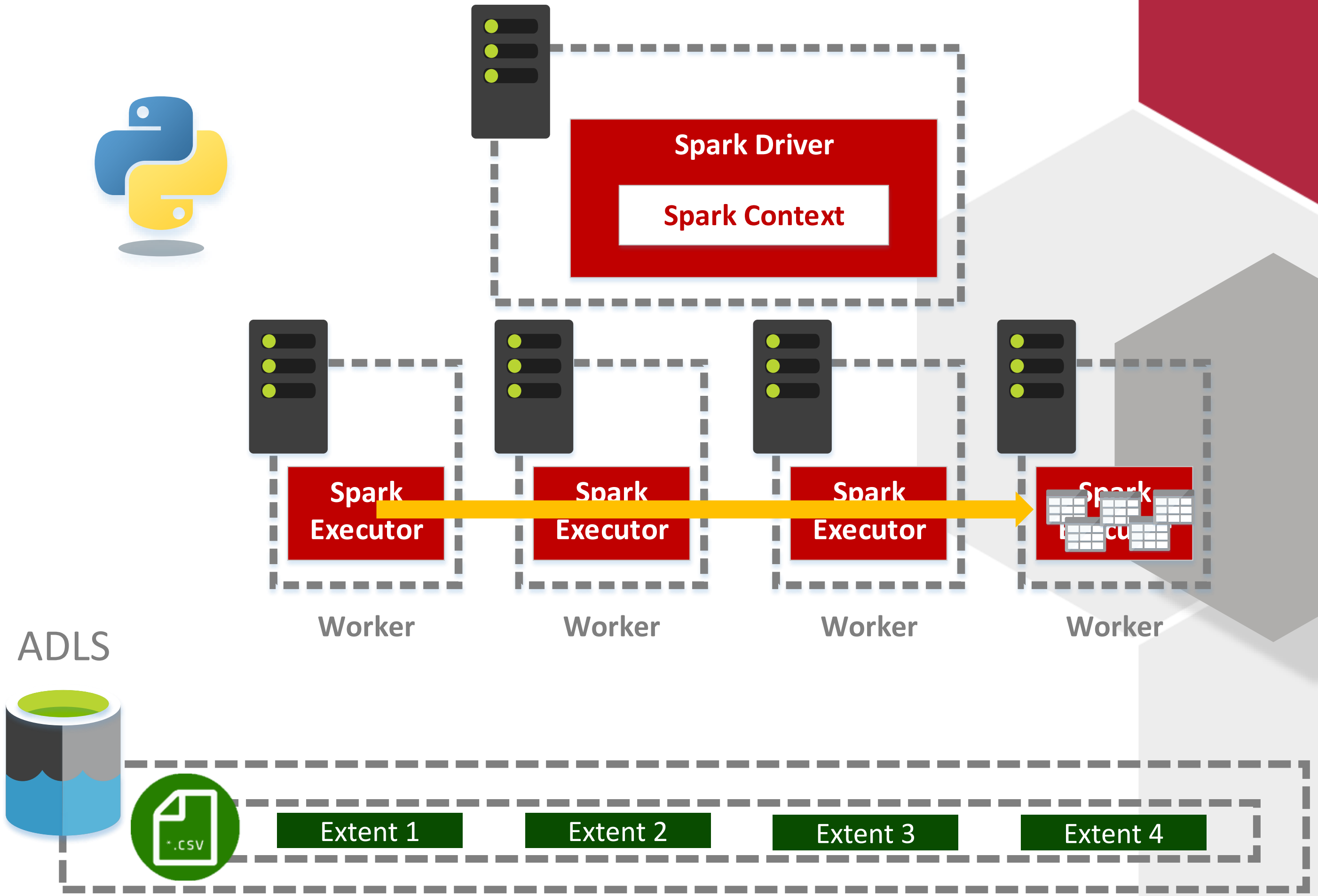


DISTRIBUTED COMPUTE



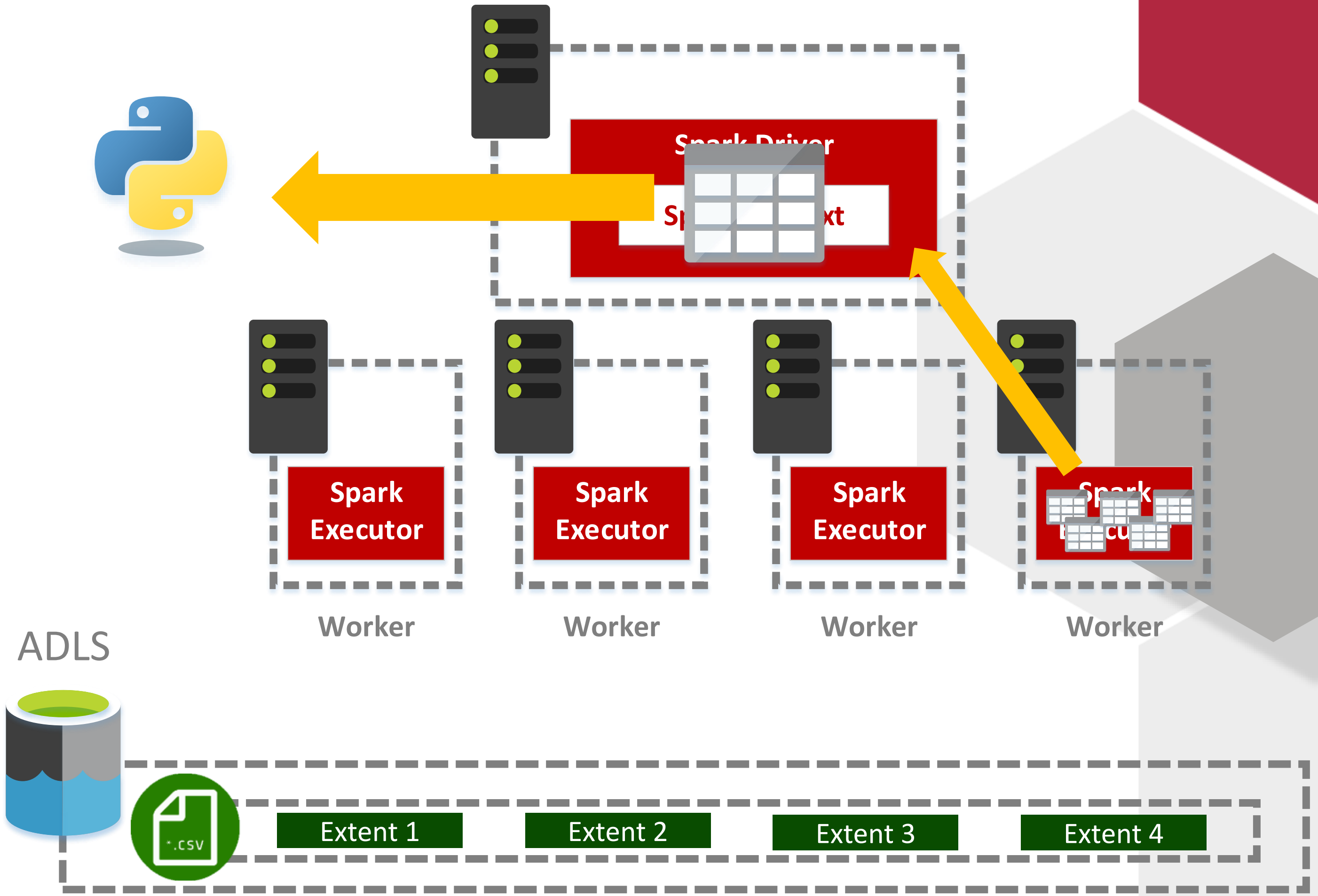
ADVANCING ANALYTICS

DISTRIBUTED COMPUTE



ADVANCING
ANALYTICS

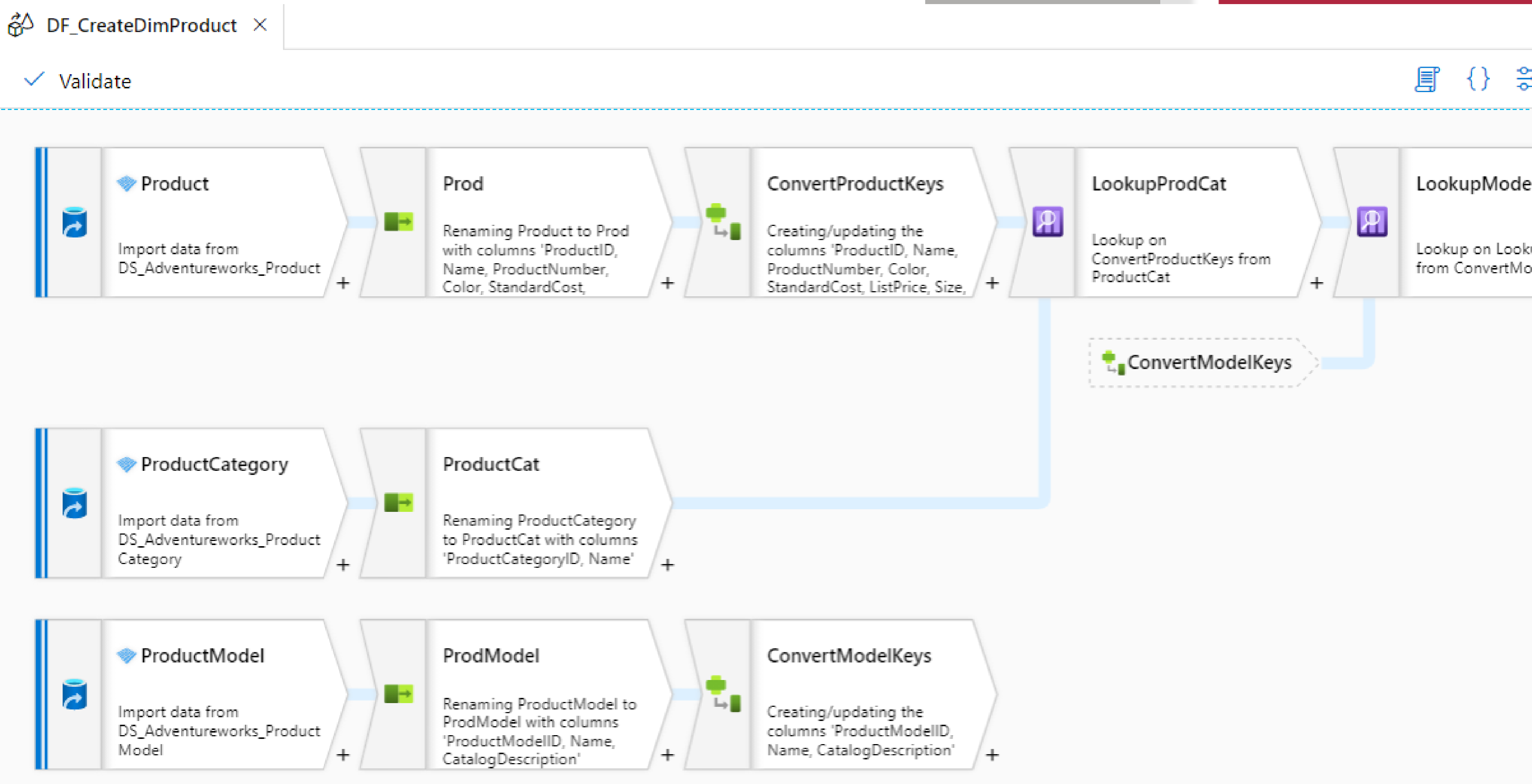
DISTRIBUTED COMPUTE



ADVANCING ANALYTICS

SPARK OPTIONS IN AZURE

AZURE DATA FACTORY – MAPPING DATA FLOWS



Product

Import data from DS_Adventureworks_Product

Prod

Renaming Product to Prod with columns 'ProductID, Name, ProductNumber, Color, StandardCost,

ConvertProductKeys

Creating/updating the columns 'ProductID, Name, ProductNumber, Color, StandardCost, ListPrice, Size,

LookupProdCat

Lookup on ConvertProductKeys from ProductCat

LookupMode

Lookup on Look from ConvertMo

ConvertModelKeys

ProductCategory

Import data from DS_Adventureworks_ProductCategory

ProductCat

Renaming ProductCategory to ProductCat with columns 'ProductCategoryID, Name'

ProductModel

Import data from DS_Adventureworks_ProductModel

ProdModel

Renaming ProductModel to ProdModel with columns 'ProductModelID, Name, CatalogDescription'

ConvertModelKeys

Creating/updating the columns 'ProductModelID, Name, CatalogDescription'

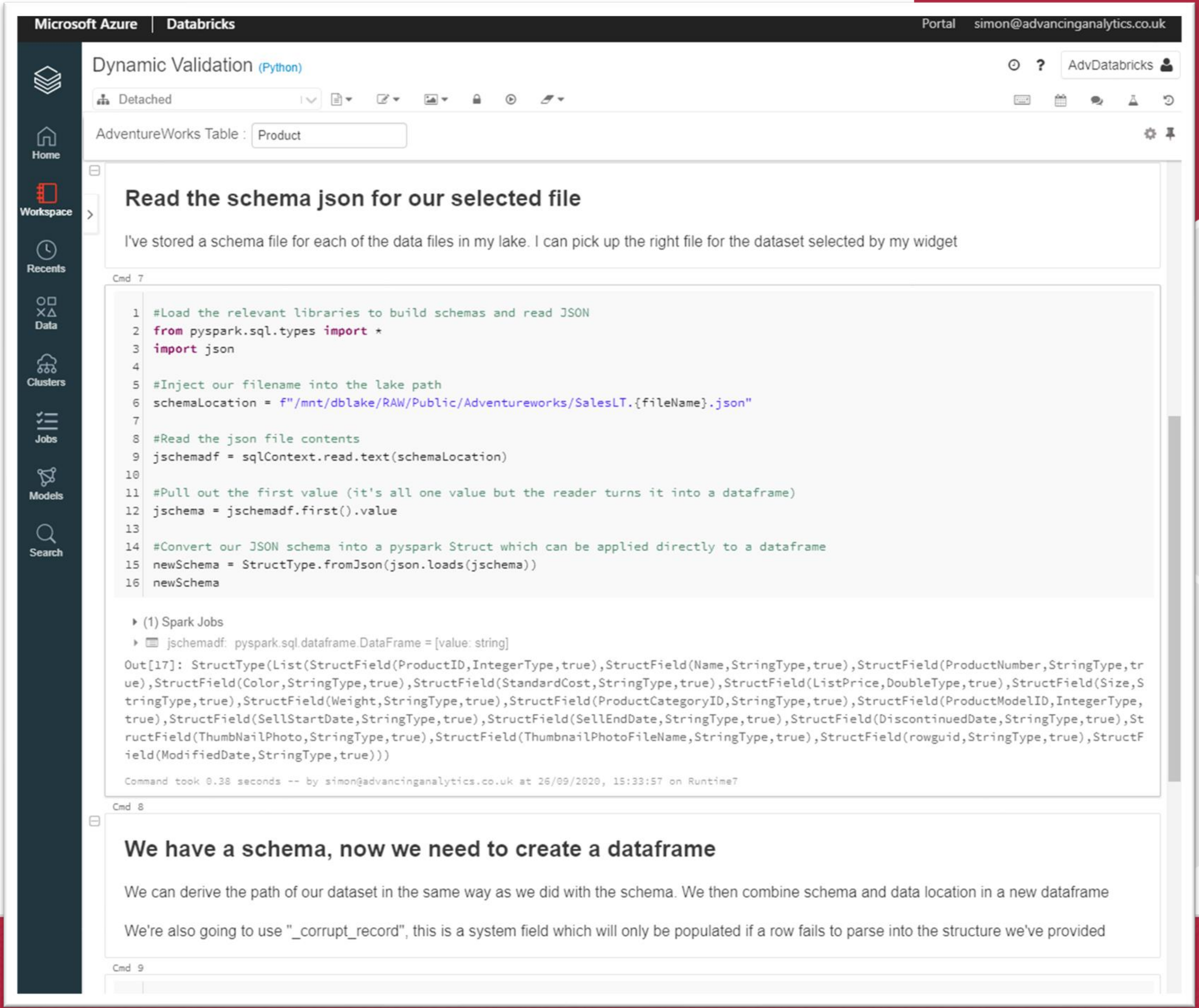
AZURE DATA FACTORY – MAPPING DATA FLOWS

Mapping Data Flows are a **GUI-Based transformation tool**, modelled on SSIS Data Flows

They use a **managed Spark Engine** which gives you the power and scale of Spark, but without the programming overhead

They are **not as flexible** as full spark but are getting better all the time!

AZURE DATABRICKS


Microsoft Azure | Databricks
Portal
simon@advancinganalytics.co.uk
Dynamic Validation (Python)
AdvDatabricks
Detached
Home
Workspace
Recents
Data
Clusters
Jobs
Models
Search

AdventureWorks Table : Product

Read the schema json for our selected file

I've stored a schema file for each of the data files in my lake. I can pick up the right file for the dataset selected by my widget

Cmd 7

```python
#Load the relevant libraries to build schemas and read JSON
from pyspark.sql.types import *
import json

#Inject our filename into the lake path
schemaLocation = f"/mnt/dblake/RAW/Public/Adventureworks/SalesLT.{fileName}.json"

#Read the json file contents
jschemadf = sqlContext.read.text(schemaLocation)

#Pull out the first value (it's all one value but the reader turns it into a dataframe)
jschema = jschemadf.first().value

#Convert our JSON schema into a pyspark Struct which can be applied directly to a dataframe
newSchema = StructType.fromJson(json.loads(jschema))
newSchema
```

(1) Spark Jobs

jschemadf: pyspark.sql.dataframe.DataFrame = [value: string]

```
Out[17]: StructType(List(StructField(ProductID,IntegerType,true),StructField(Name,StringType,true),StructField(ProductNumber,StringType,tr
ue),StructField(Color,StringType,true),StructField(StandardCost,StringType,true),StructField(ListPrice,DoubleType,true),StructField(Size,S
tringType,true),StructField(Weight,StringType,true),StructField(ProductCategoryID,StringType,true),StructField(ProductModelID,IntegerType,
true),StructField(SellStartDate,StringType,true),StructField(SellEndDate,StringType,true),StructField(DiscontinuedDate,StringType,true),St
ructField(ThumbNailPhoto,StringType,true),StructField(ThumbnailPhotoFileName,StringType,true),StructField(rowguid,StringType,true),StructF
ield(ModifiedDate,StringType,true)))
```

Command took 0.38 seconds -- by simon@advancinganalytics.co.uk at 26/09/2020, 15:33:57 on Runtime7

Cmd 8

We have a schema, now we need to create a dataframe

We can derive the path of our dataset in the same way as we did with the schema. We then combine schema and data location in a new dataframe

We're also going to use "_corrupt_record", this is a system field which will only be populated if a row fails to parse into the structure we've provided

Cmd 9

ADVANCING ANALYTICS

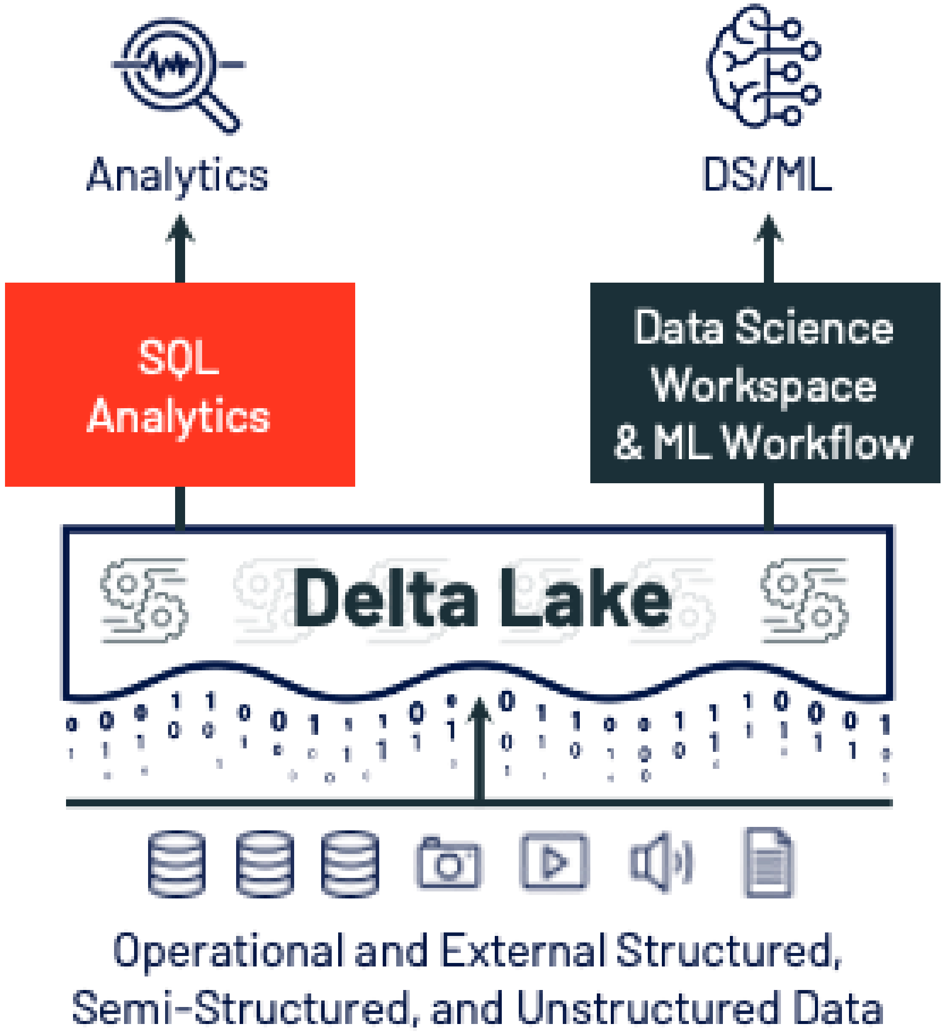
AZURE DATABRICKS

Databricks is a third-party company, founded by the **team who invented spark**

They provide an **Azure-native, managed Spark platform**

Databricks will generally have the **most advanced spark engine** and maintain a fast release cadence

DATABRICKS SQL ANALYTICS


Analytics
DS/ML
SQL Analytics
Data Science Workspace & ML Workflow
Delta Lake
Operational and External Structured, Semi-Structured, and Unstructured Data

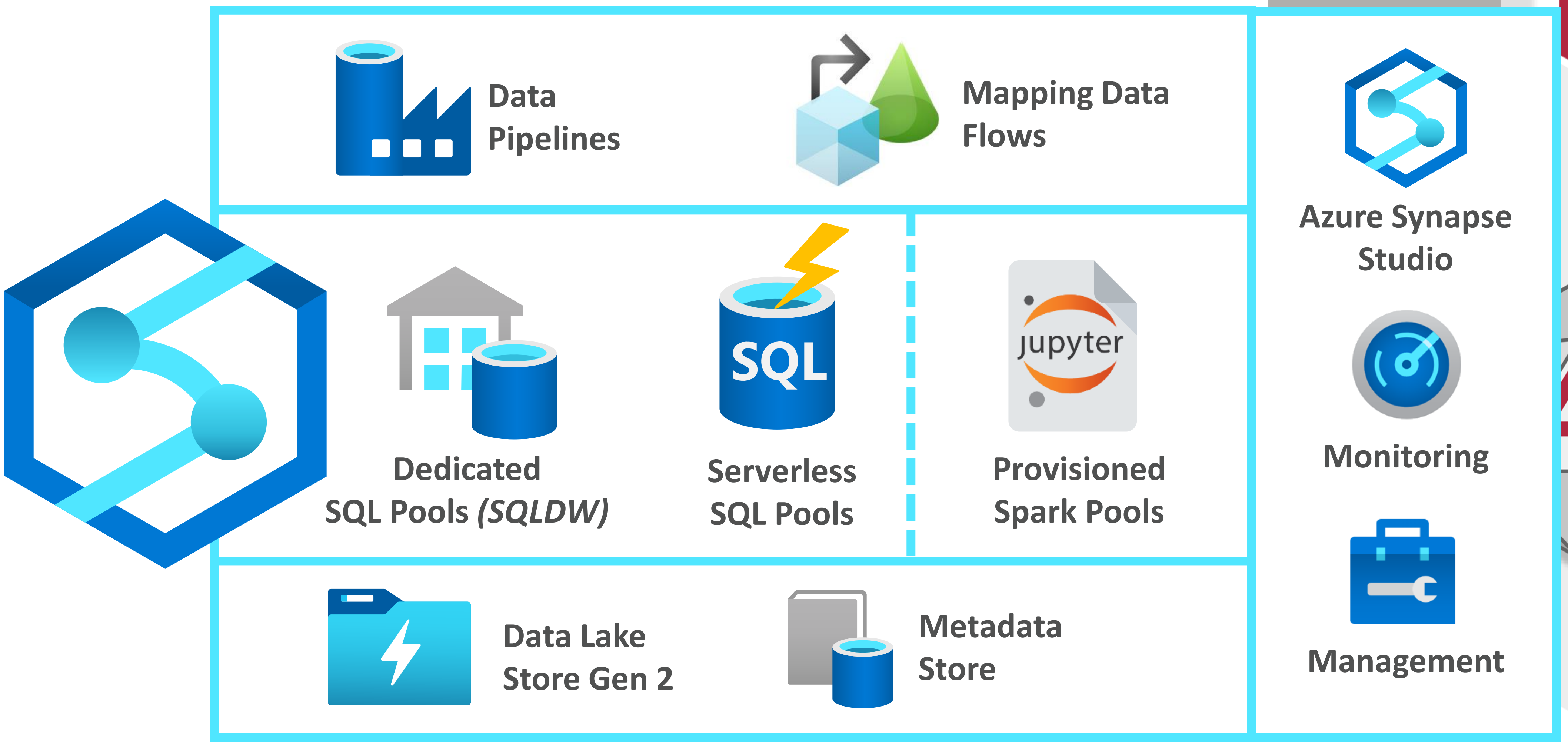
AZURE DATABRICKS

- The Databricks Workspace
- Clusters
- Notebooks

AZURE SYNAPSE ANALYTICS

Where to start…

AZURE SYNAPSE ANALYTICS



Data Pipelines

Mapping Data Flows

Dedicated SQL Pools *(SQLDW)*

Serverless SQL Pools

Provisioned Spark Pools

Data Lake Store Gen 2

Metadata Store

Azure Synapse Studio

Monitoring

Management

AZURE SYNAPSE ANALYTICS - SPARK POOLS

- Billed Per Session Uptime
- Scala, Python, C#, SQL
- Dynamic Workflows, Machine Learning & Unusual Data Types

- Session management is… interesting

SERVERLESS SQL POOLS

- Billed Per TB Read
- T-SQL
- Ad-hoc/Occasional access

- Unpredictable Billing
- Very Black-box (but is that bad?)

DEDICATED SQL POOLS

- Billed Per Hour
- T-SQL
- Huge Datasets & Formal Modelling

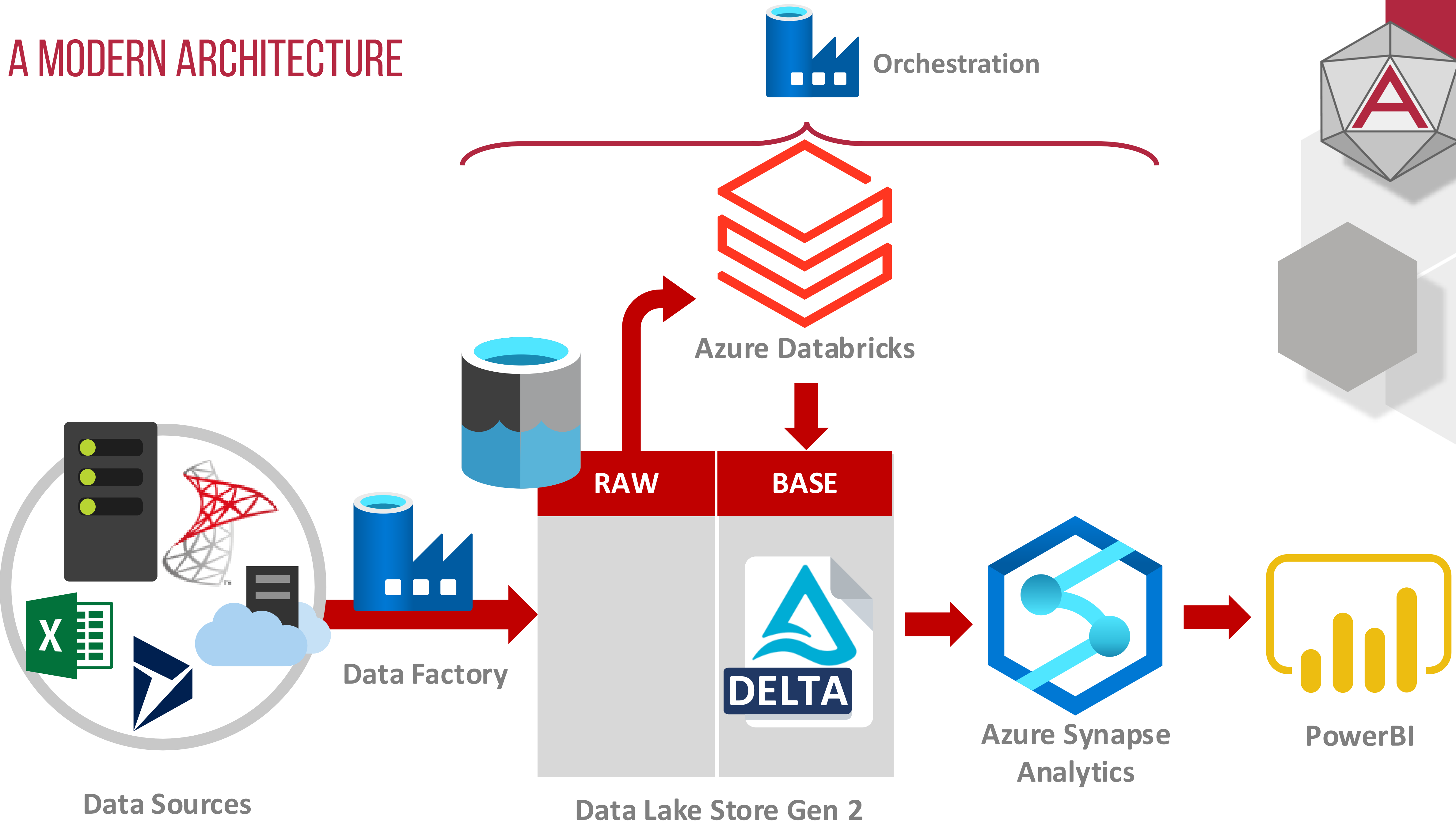
- Inflexible Scaling
- Can be complex to distribute tables

AZURE SYNAPSE ANALYTICS

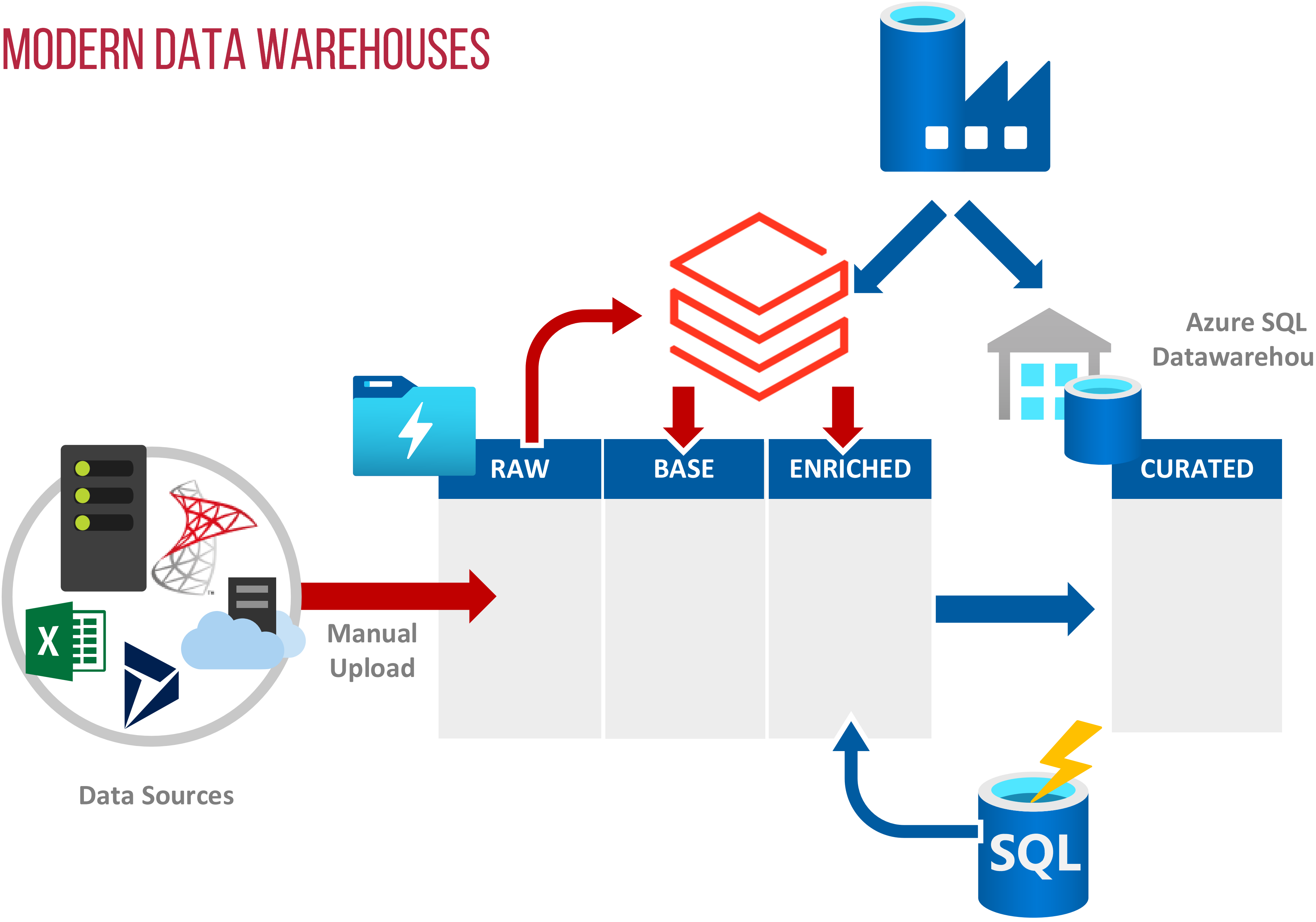
- Synapse Workspace Overview
- Serverless SQL
- Spark Pools

SO WHERE DOES THAT LEAVE US?

A MODERN ARCHITECTURE



MODERN DATA WAREHOUSES



Azure SQL
Datawarehouse
RAW
BASE
ENRICHED
CURATED
Manual
Upload
Data Sources
SQL