

Pipelines and Packages: Introduction to Azure Data Factory



Cathrine Wilhelmsen

DATA:Scotland · September 13th, 2019



DATA:Scotland



Microsoft



Arnold Clark



SentryOne®



Pipelines and Packages: Introduction to Azure Data Factory

As Data Engineers and ETL Developers, our main responsibilities are to move, transform, integrate and prepare data for our end users as quickly and efficiently as possible. With the ever-increasing volume and variety of data, this can easily start to feel like a daunting task.

Azure Data Factory (ADF) is a hybrid data integration service that lets you build, orchestrate and monitor complex and scalable data pipelines - without writing any code. The first version of Azure Data Factory may not have lived entirely up to its nickname "SSIS in the Cloud", but the second version has been drastically improved and expanded with new capabilities.

But wait, what's that? You have already invested years and millions in a comprehensive SSIS solution, you say? No problem! You can lift and shift your existing SSIS packages into Azure Data Factory to start modernizing your solution while retaining the investments you have already made.

In this session, we will first go through the fundamentals of Azure Data Factory and see how easy it is to build new data pipelines or migrate your existing SSIS packages. Then, we will explore some of the major improvements in Azure Data Factory v2, including the new Mapping Data Flows. Finally, we will look at design patterns and best practices for development to speed up productivity while keeping costs down.

cathrine

WILHELMSEN



@cathrinew



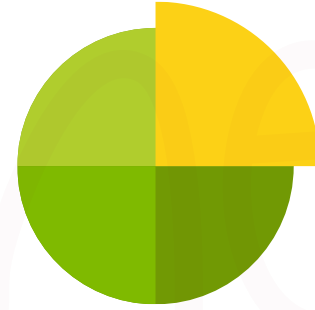
cathrinew.net



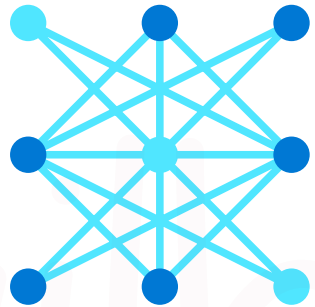
Data Warehousing



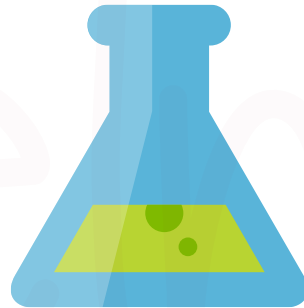
Big Data and Analytics



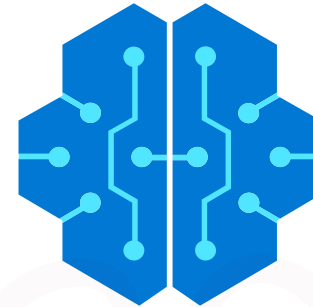
Business Intelligence



Artificial Intelligence



Data Science



Machine Learning



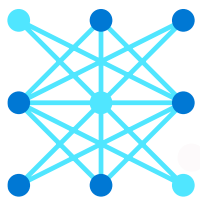
Big Data and Analytics



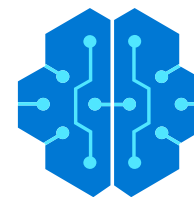
Data Warehousing



Business Intelligence



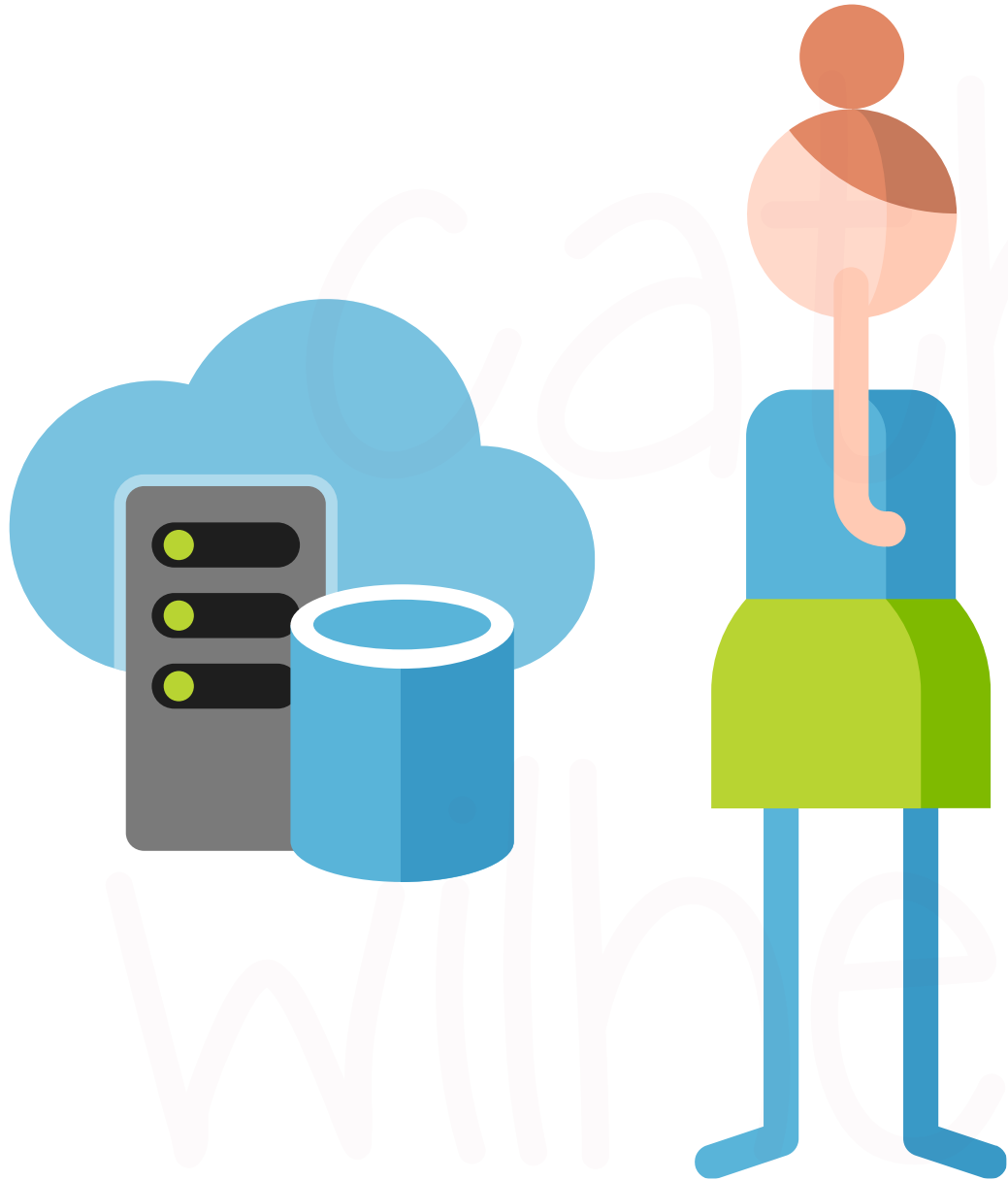
Artificial Intelligence



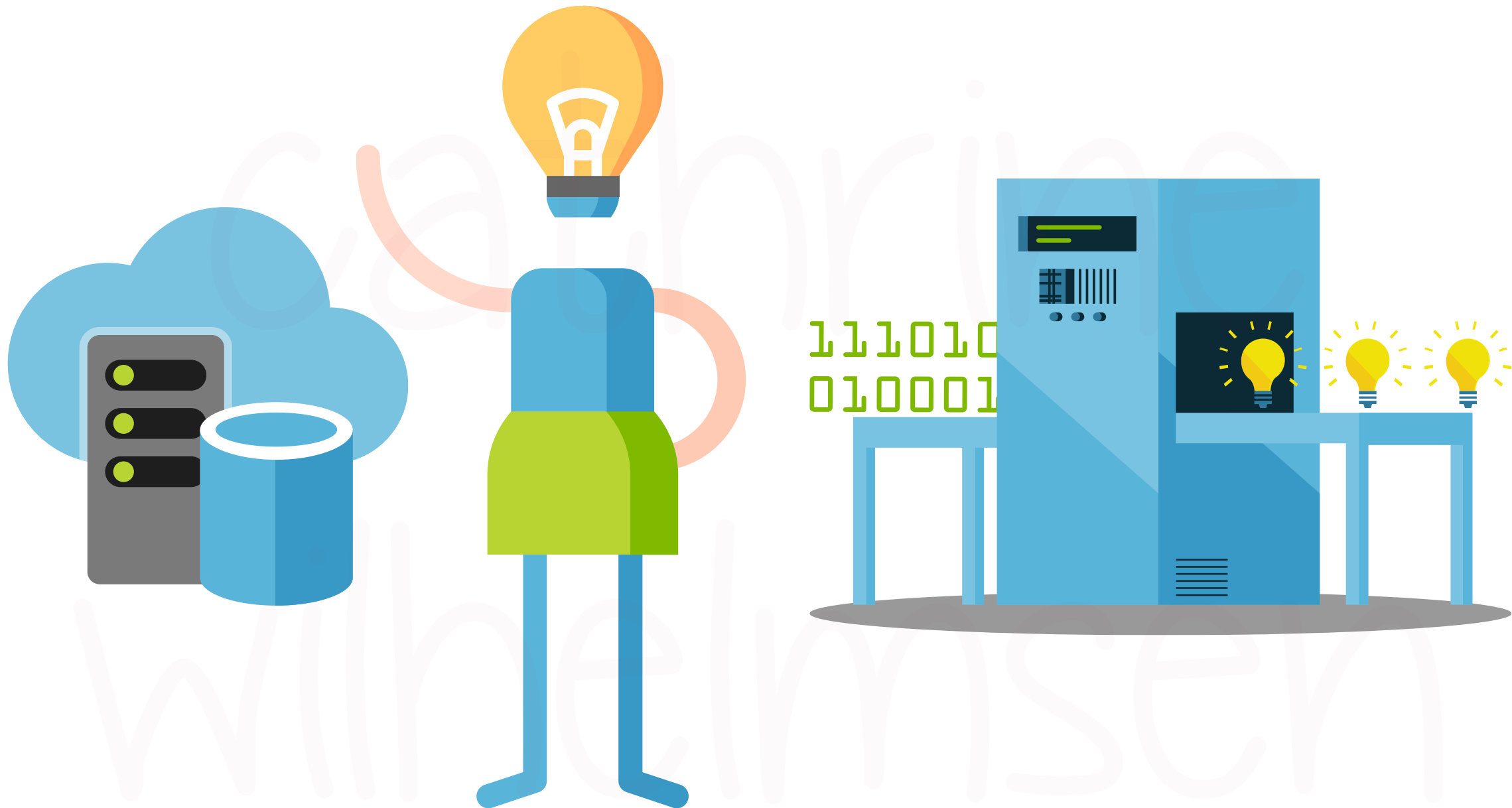
Machine Learning



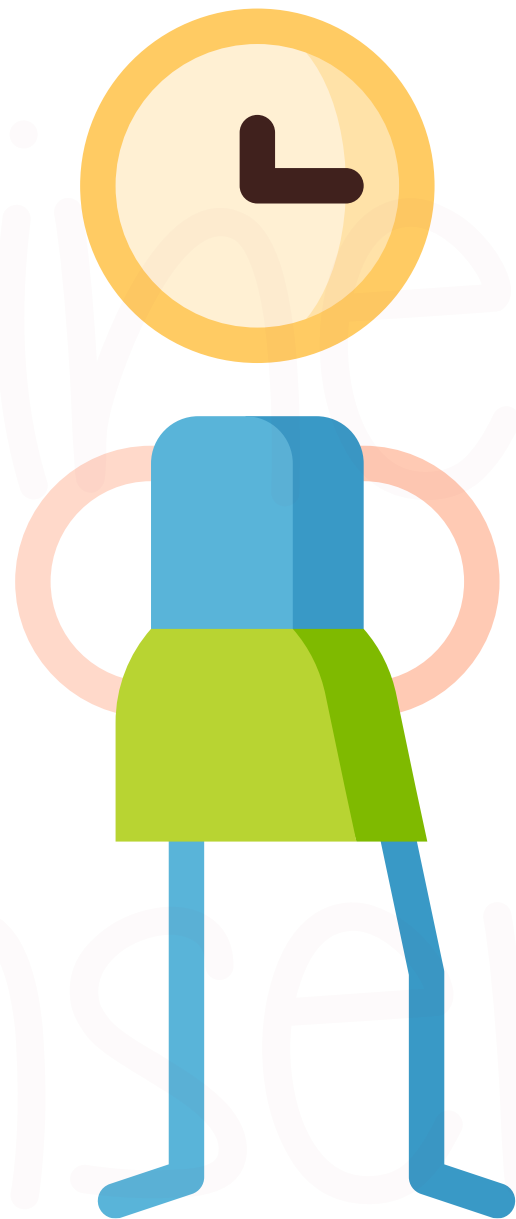
Data Science

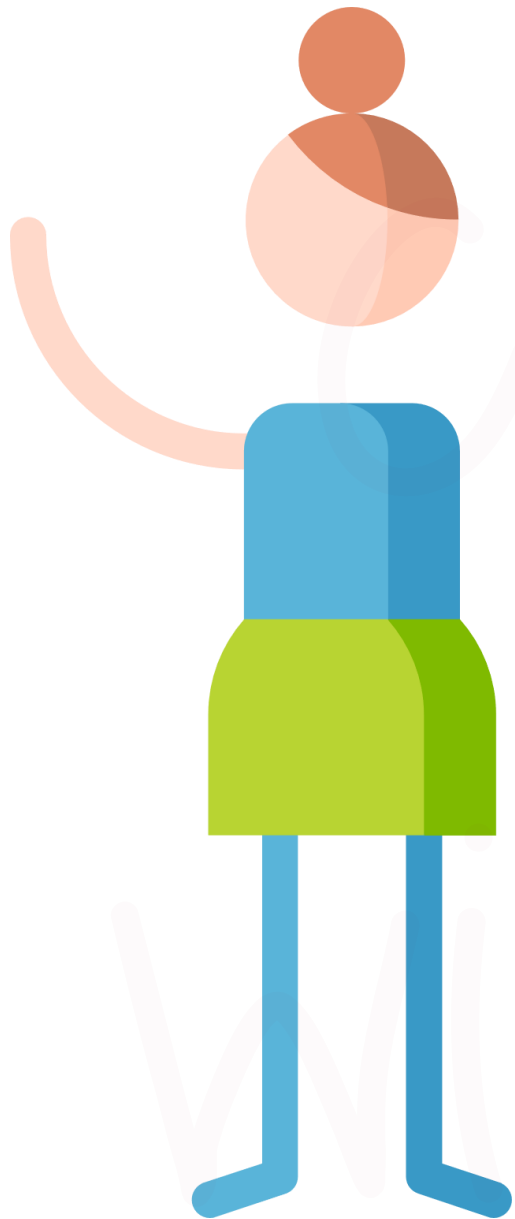


Collect
Store
Transform
Integrate
Prepare



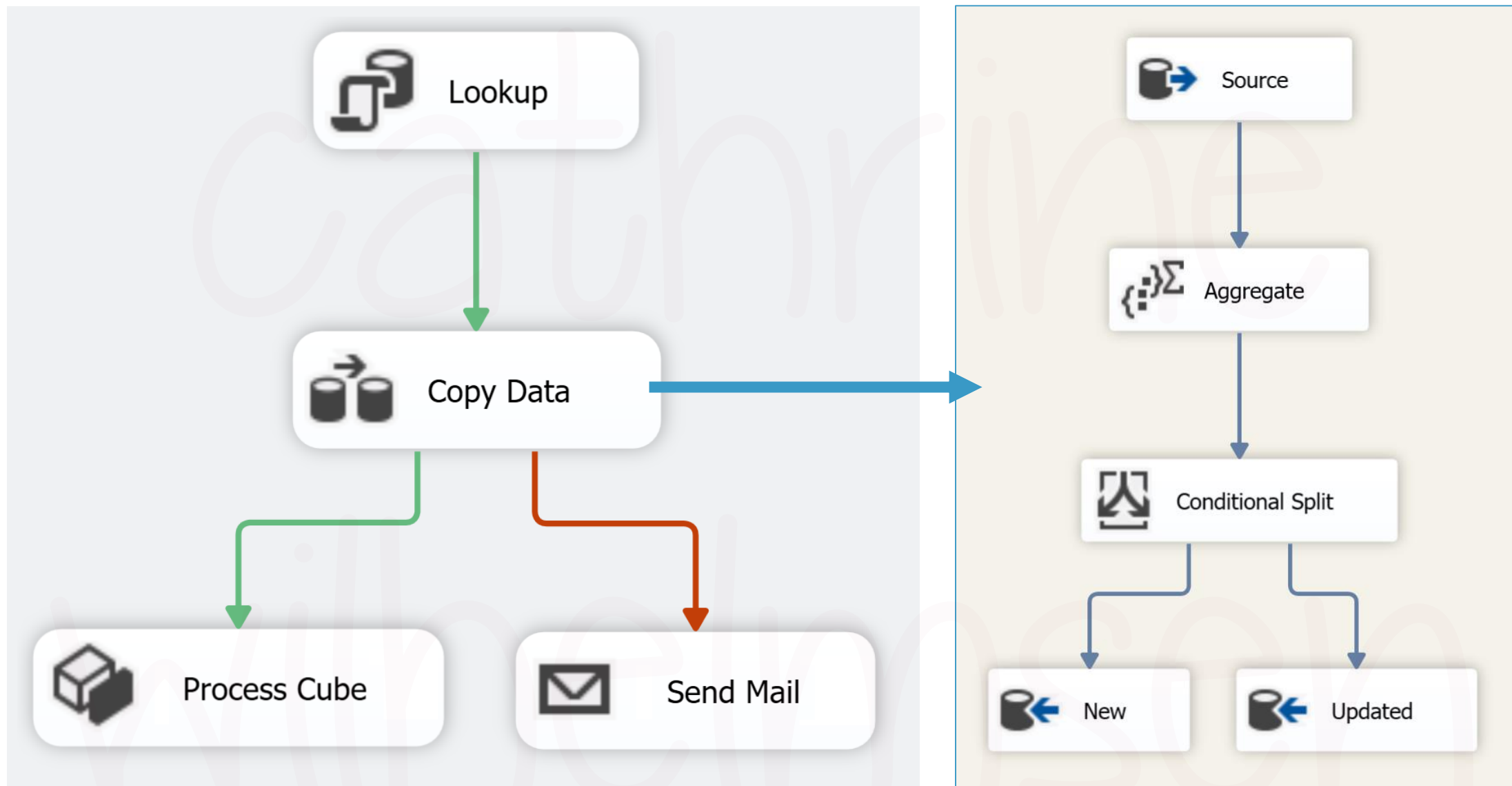
10 years ago...



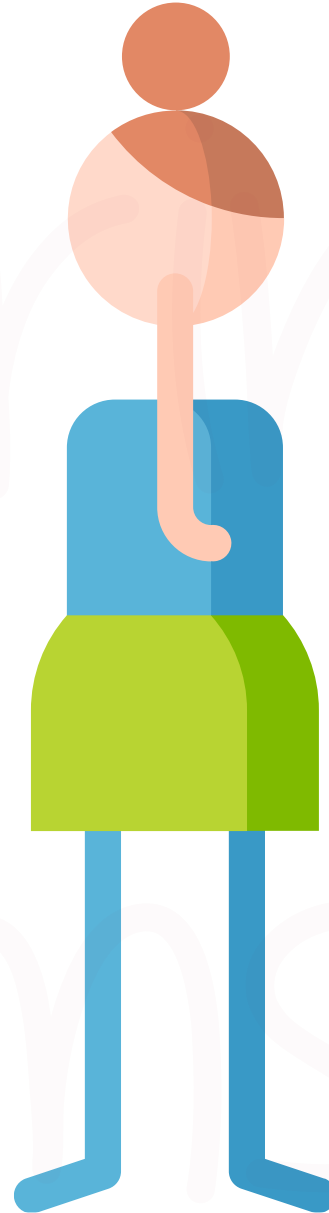


SSIS

SQL Server Integration Services



Then...

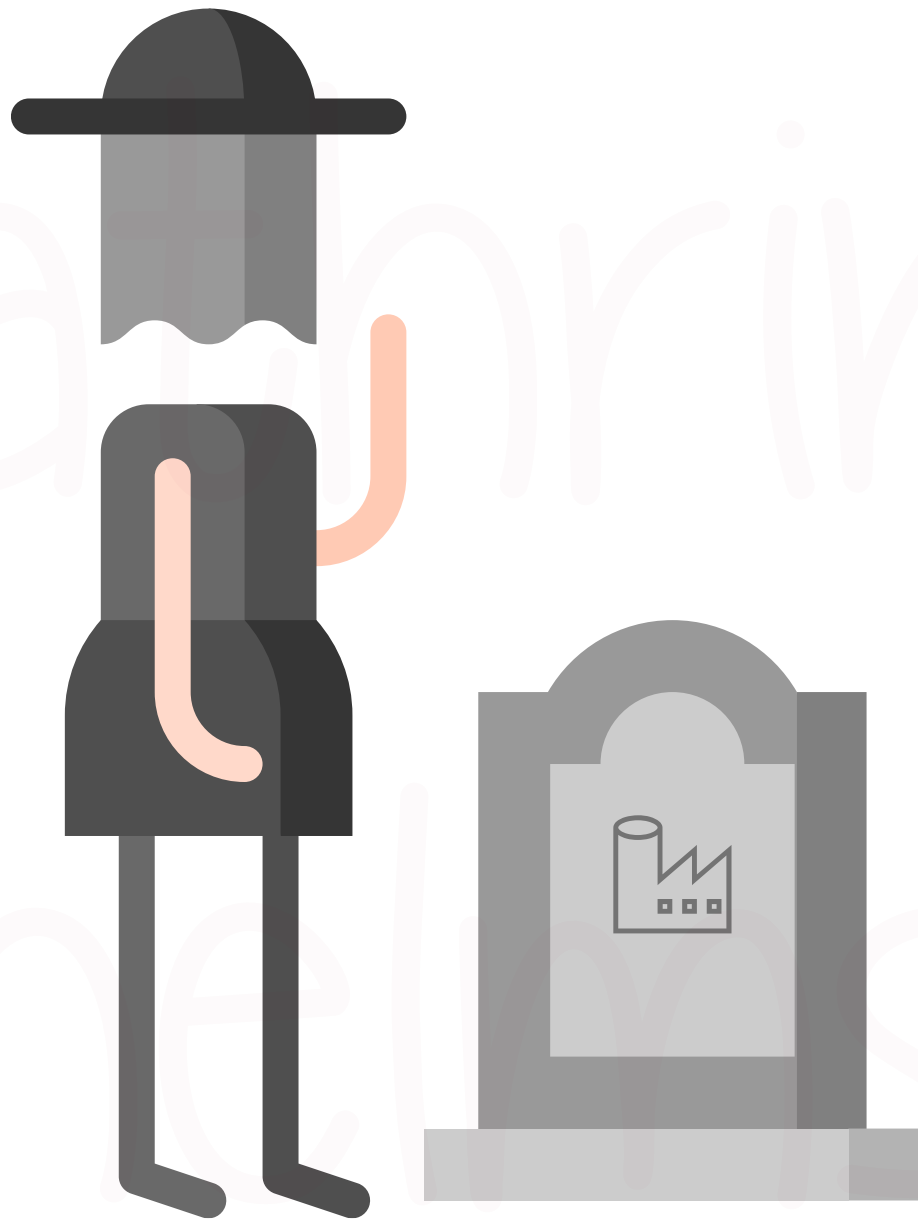


ADF v1

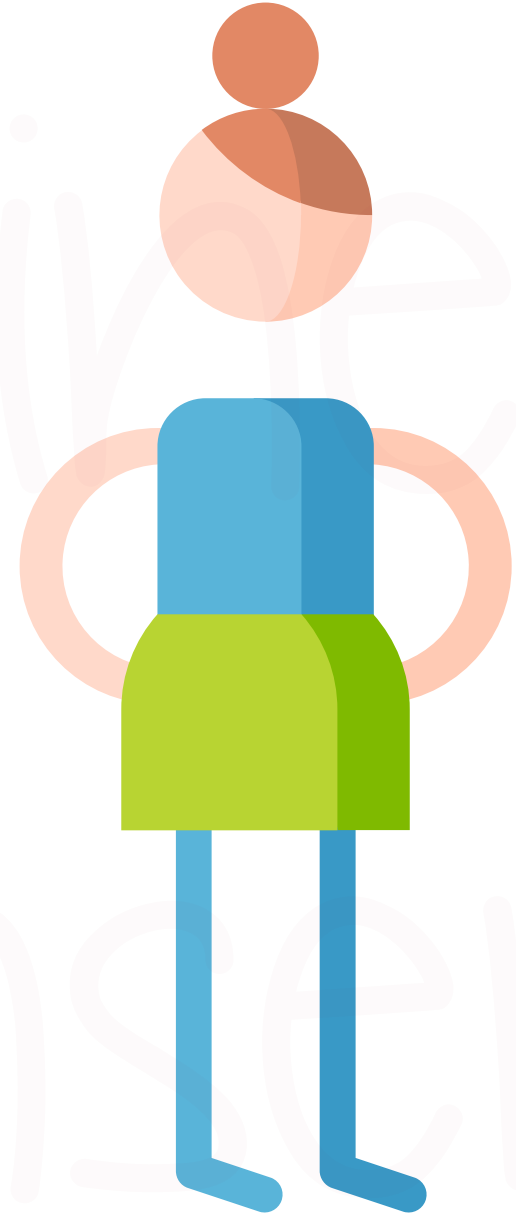
Azure Data Factory Version 1

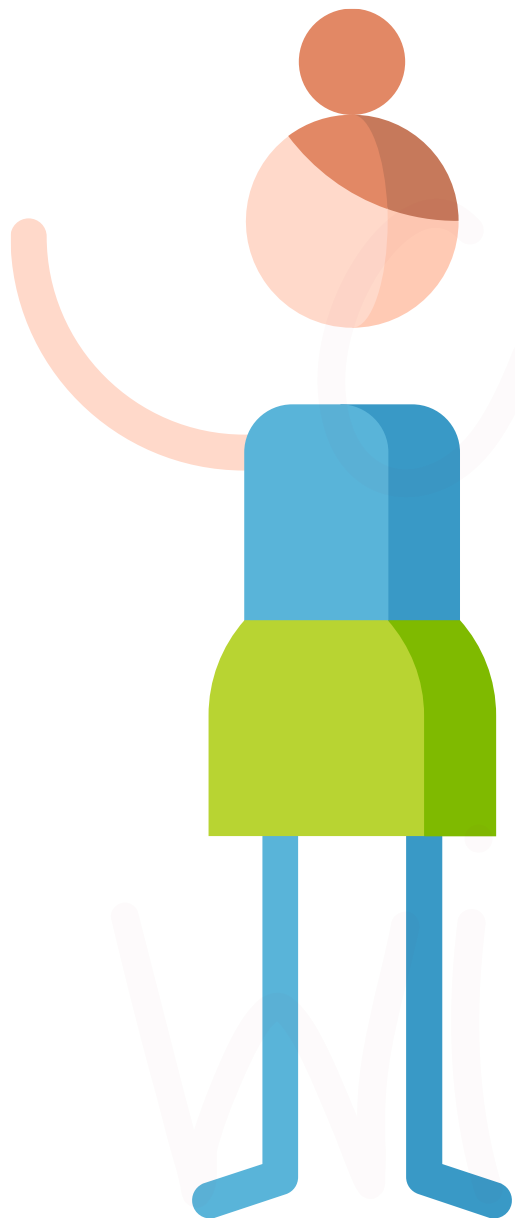






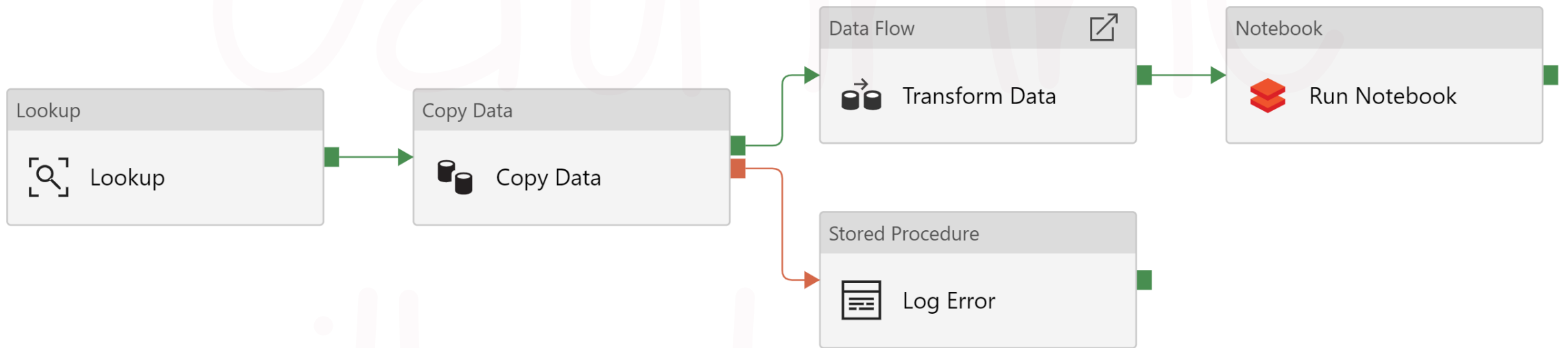
Today...





ADF v2

Azure Data Factory Version 2



Why?

Stop using SSIS?

Move to ADF?

How?

What?

Existing solution?





Azure

Data Factory

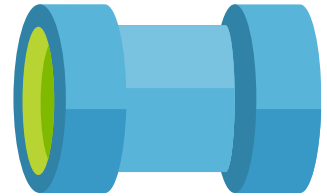
What is Azure Data Factory?



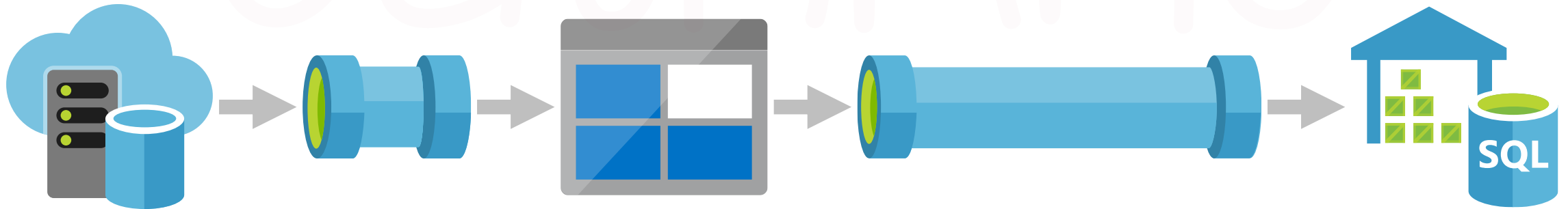
Hybrid data integration service

Complex and scalable pipelines

No-code ETL/ELT data flows



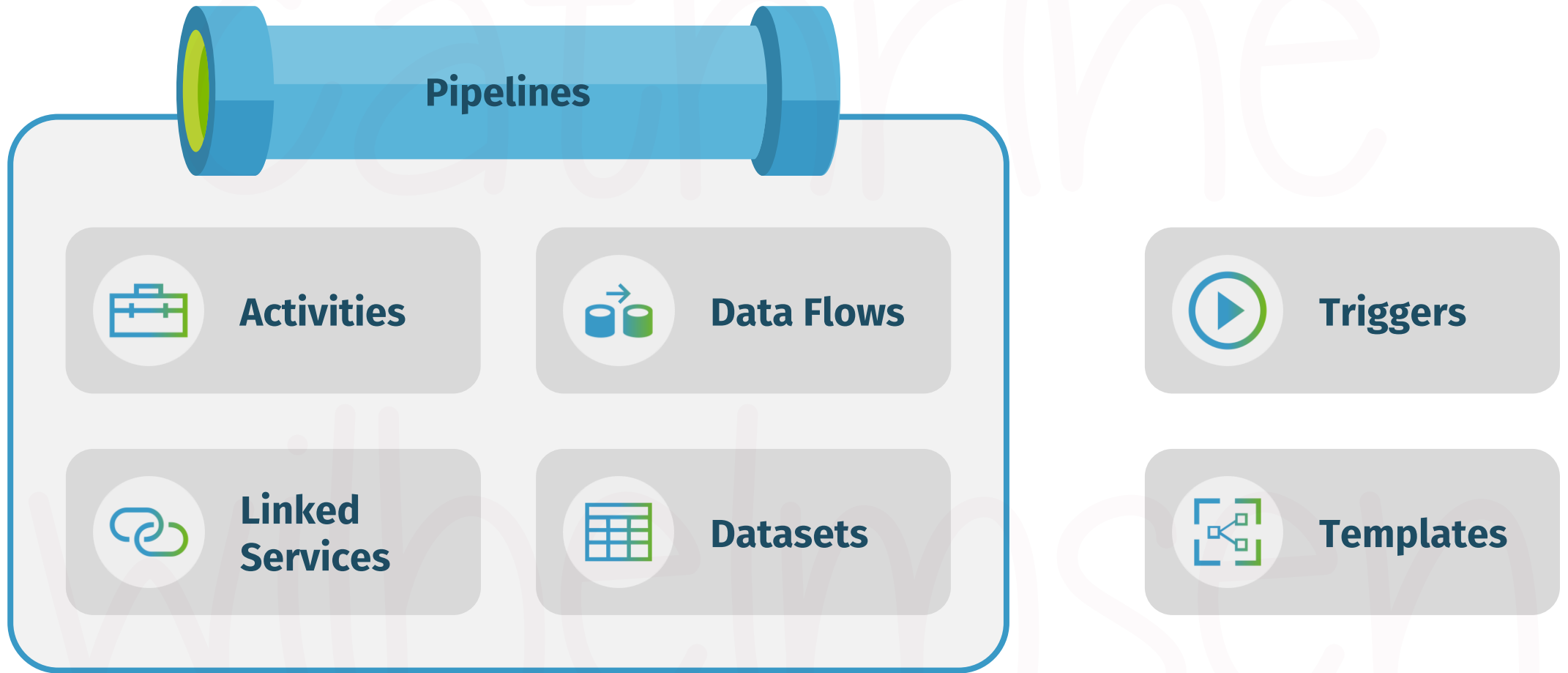
What can you do in Azure Data Factory?



Copy Data

Transform Data

What is inside Azure Data Factory?



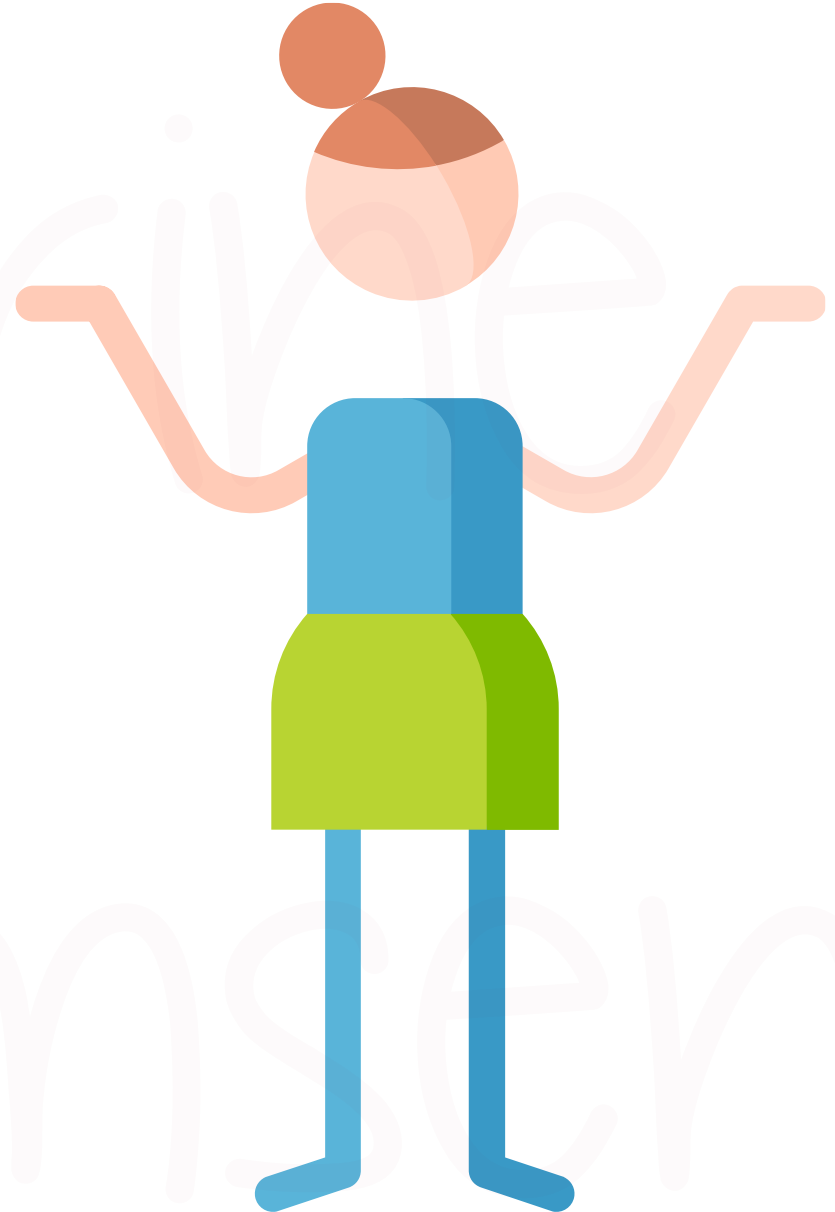
DEMO

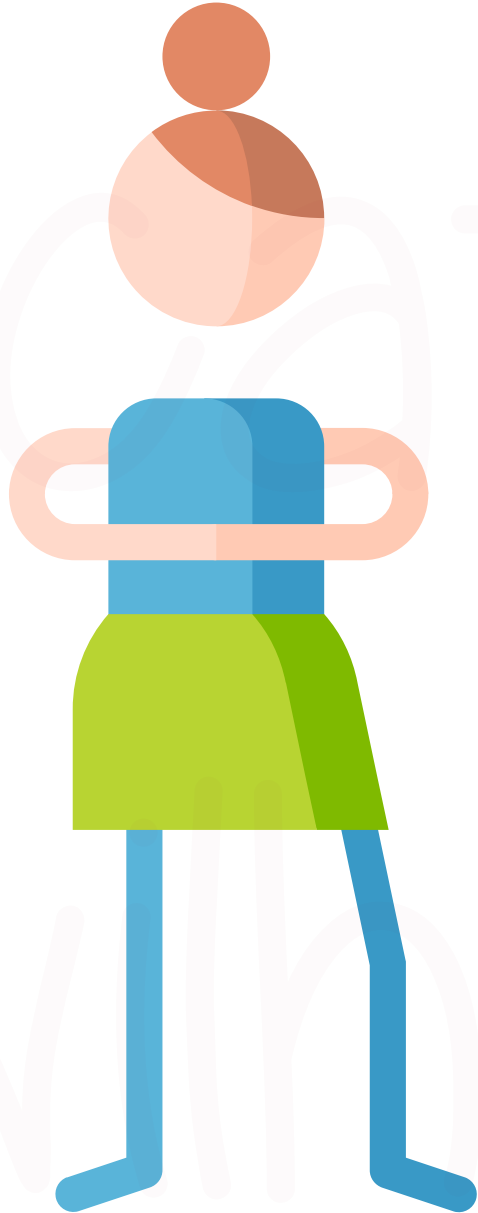
Let's look inside Azure Data Factory!



Wait...

I already have
thousands of SSIS
packages!





And...

You told me to use
this Biml thing!



SSIS

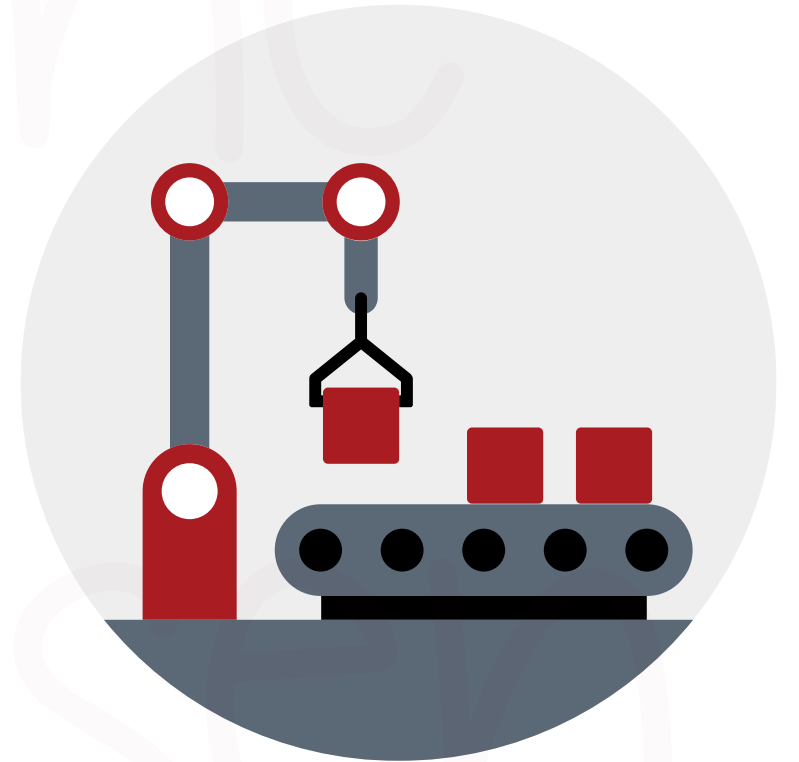
Lift and Shift

What does Lift and Shift mean?



Lift up existing SSIS packages

Shift them to a new location



Why should you Lift and Shift SSIS?



Modernize while retaining investments



Reduce maintenance and costs (*)



Continue to use familiar tools



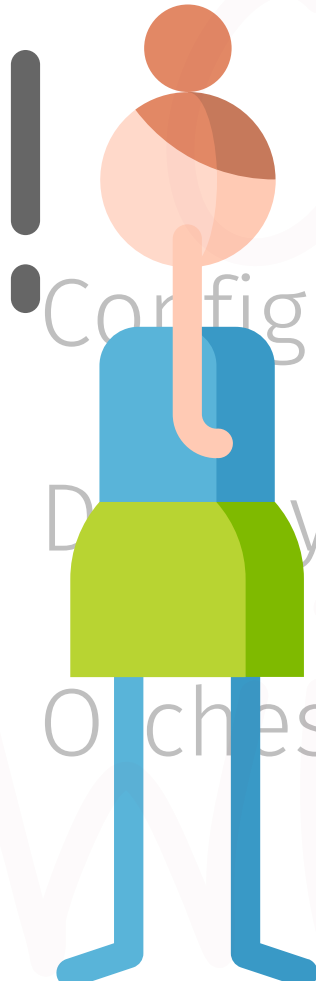
How do you Lift and Shift SSIS?



- 1.** Configure Azure-SSIS Integration Runtime
- 2.** Deploy SSIS Packages to SSISDB in Azure
- 3.** Orchestrate SSIS Packages in Azure Data Factory

How do you Lift and Shift SSIS?



- 
1. Configure Azure-SSIS Integration Runtime
 2. Deploy SSIS Packages to SSISDB in Azure
 3. Orchestrate SSIS Packages in Azure Data Factory

Azure-SSIS Integration Runtime



Managed cluster of Azure VMs dedicated to SSIS

Billed while running (*like all VMs*)

Manage cost by running when necessary



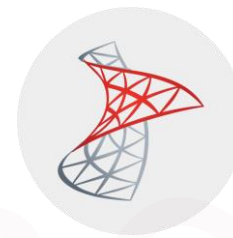
DEMO

**Let's lift and shift
some SSIS Packages!**





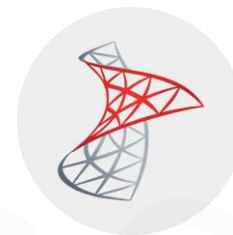
ADF vs SSIS



Pipeline	≈	Package
Linked Service	≈	Connection Manager
Source	≈	Source
Sink	≈	Destination
Activity	≈	Control Flow Task
Data Flow	≈	Data Flow



ADF vs SSIS



Pipeline	≈	Package
Linked Service	≈	Connection Manager
Source	≈	Source
Sink	≈	Destination
Activity	≈	Control Flow Task
Data Flow	≈	Data Flow



Mapping Data Flows

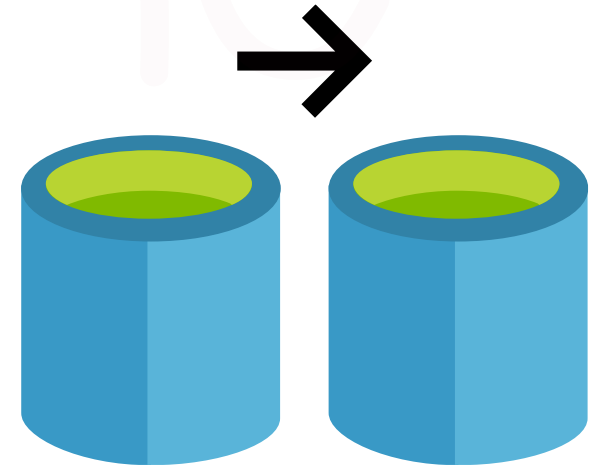
What are Mapping Data Flows?



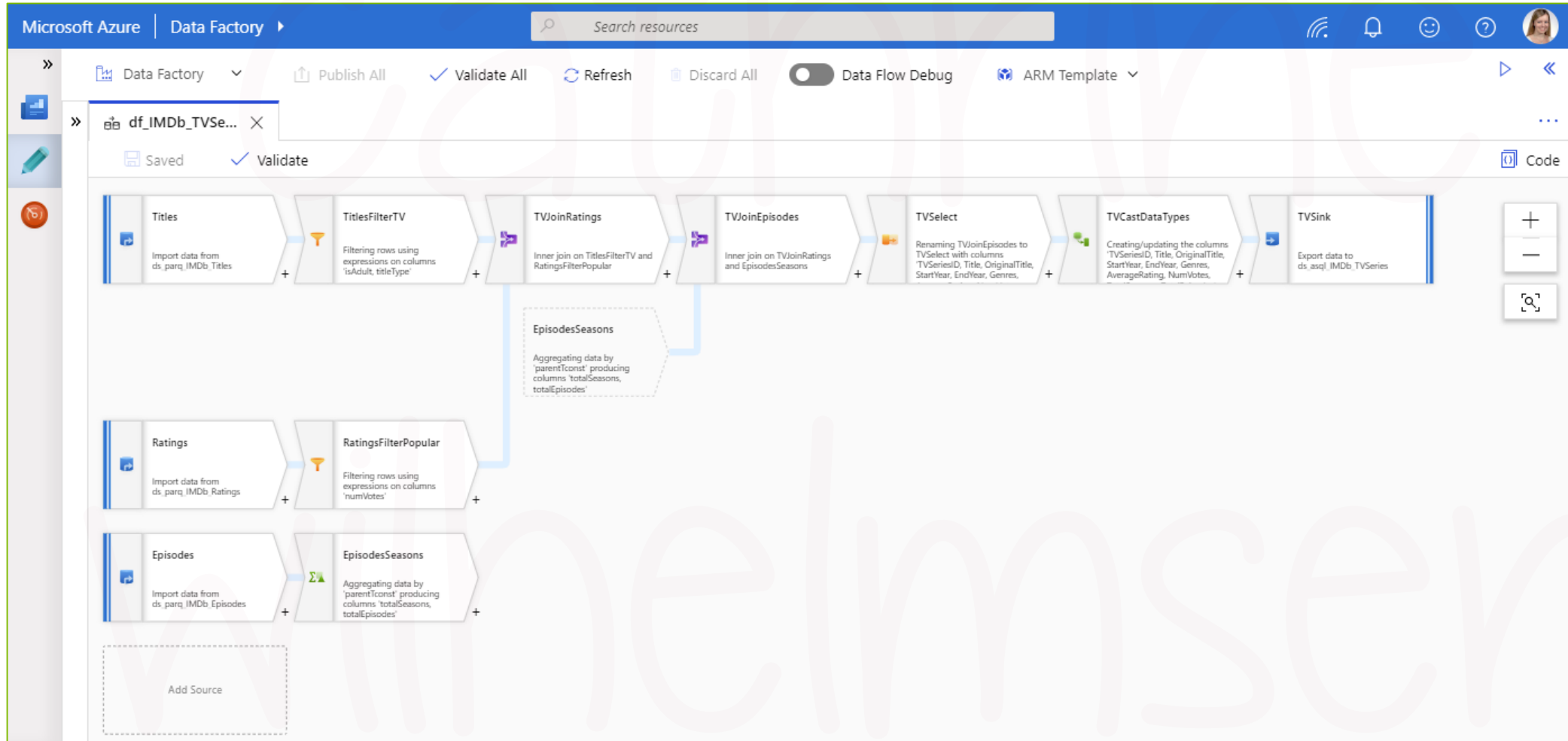
Data transformation at scale

Runs on Azure Databricks

Visual editor, no-code experience



How do Mapping Data Flows work?



Why use Mapping Data Flows?



Transform Data

Upsert Data

Load a Data Warehouse

Handle schema drift

What is Schema Drift?

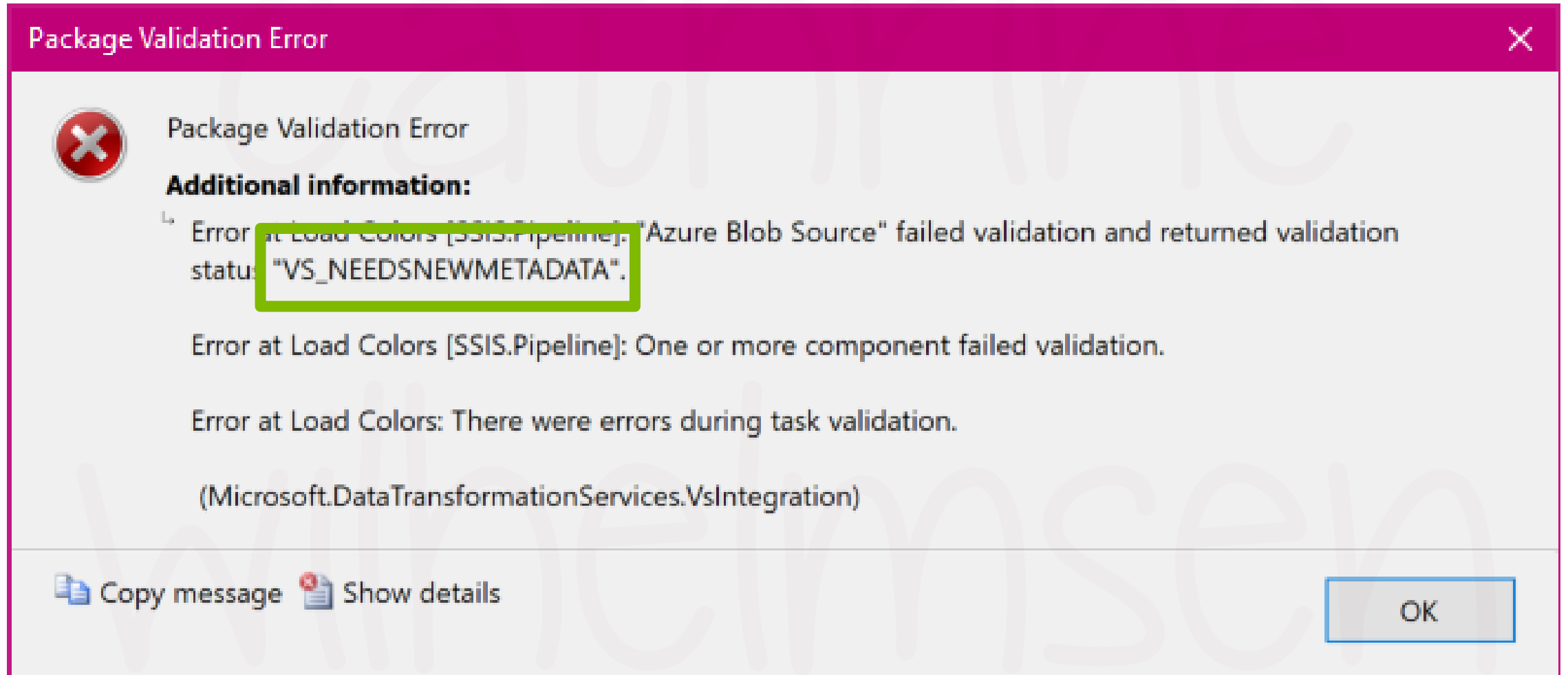


Rapidly changing source files and metadata:

- Added / Removed Columns
- Renamed Column Names
- Changed Data Types

If not handled properly, Schema Drift can (*and most likely will*) cause problems in the upstream pipeline

Schema Drift in SSIS



Schema Drift in ADF

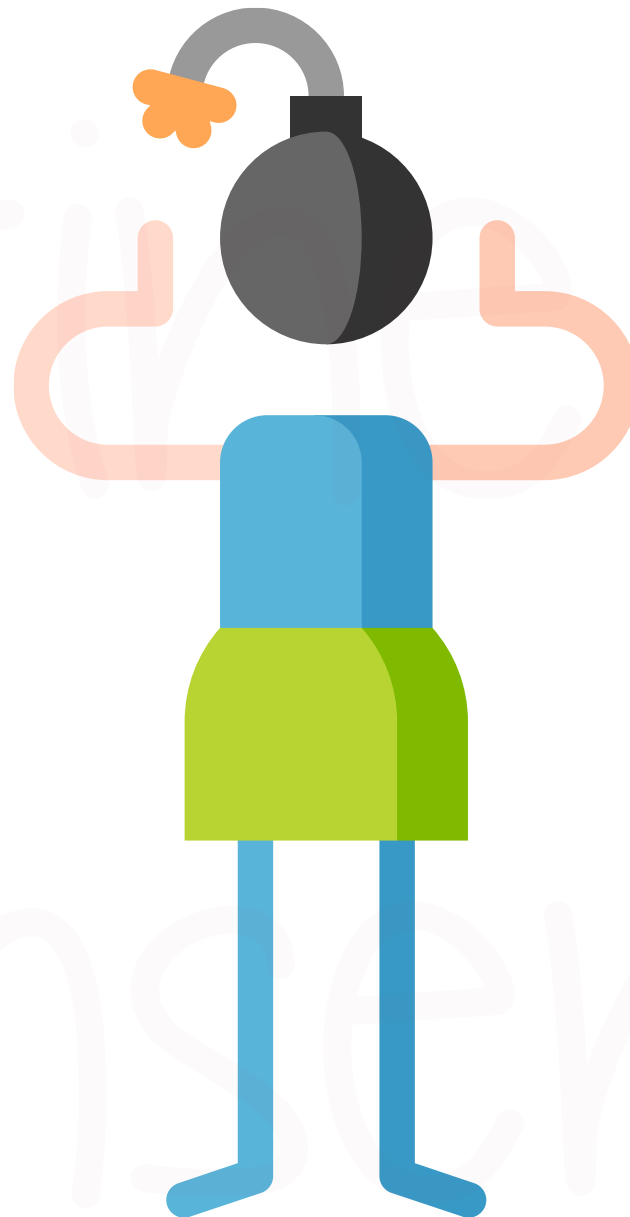
The screenshot shows an Azure Data Factory (ADF) pipeline interface. At the top, a 'Copy Data' activity named 'Colors - Blob to ASQL' is displayed with a red warning icon. Below the activity, a toolbar contains icons for adding, removing, locking, zooming, and other actions. The 'Output' tab is selected, showing the 'Pipeline Run ID: 3fb3e768-1162-4d99-bac0-218762cc45c'. A table lists the activity 'Colors - Blob to ASQL' with a status of 'Failed' and a duration of '00:02:47'. An error message dialog is overlaid on the right, detailing the failure.

Error

```
{
  "errorCode": "2200",
  "message":
    "ErrorCode=UserErrorInvalidColumnName,'Type=Microsoft.Data
    Transfer.Common.Shared.HybridDeliveryException,Message=Col
    umn 'comment' does not exist in the table '[lego].[Colors]',
    SourceName=Microsoft.DataTransfer.ClientLibrary,DatabaseName=
    'Entertainment',Source=Microsoft.DataTransfer.ClientLibrary,'Ty
    pe=System.InvalidOperationException,Message=The given
    ColumnMapping does not match up with any column in the
```

NAME	TYPE	RUN START	DURATION
Colors - Blob to ASQL	Copy	00:02:47	Failed

Oh no!



DEMO

**Let's transform
some data!**



Lessons Learned

In ADF, everything has a price

SSIS best practices != ADF best practices

Learn how to learn and adapt





Good luck!

thank you!



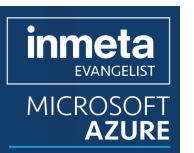
hi@cathrinew.net



@cathrinew



cathrinew.net



Session Feedback
bit.ly/2019sfeedback



Event Feedback
bit.ly/2019efeedback

DATA:Scotland