24HOURS OF PASS

Pipelines and Packages: Introduction to Azure Data Factory

Cathrine Wilhelmsen, Senior BI Consultant, Inmeta
Moderated By: Giuliana Grecco

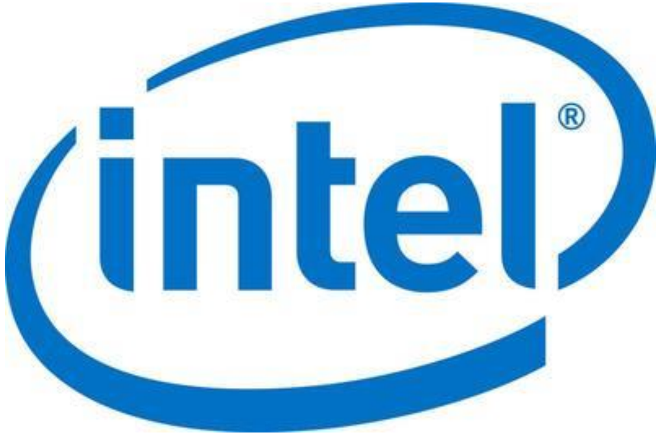
Technical Assistance

If you require assistance during the session, type your inquiry into the question pane on the right side.

Maximize your screen with the zoom button on the top of the presentation window.

Please fill in the short evaluation following the session. It will appear in your web browser.

Thank you to our Presenting Sponsors



Explore everything PASS has to offer

Free Online Resources

Newsletters

PASS.org

PASS SUMMIT

PASS' flagship event
November 5-8
Seattle, Washington

PASS LOCAL GROUPS

Local user groups
around the world

PASS SQLSATURDAY

Free 1-day local
training events

PASS VIRTUAL GROUPS

Online special interest
user groups

PASS MARATHON

Business analytics
training

PASS VOLUNTEERS

Get involved

Cathrine Wilhelmsen

Senior BI Consultant, *Inmeta*

in /cathrinewilhelmsen

@cathrinew

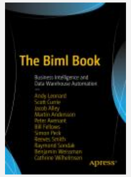
cathrinew.net

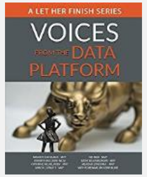
hi@cathrinew.net

Work Things

Senior BI Consultant and Microsoft Data Platform Tech Lead in Inmeta, Norway.
Specialties: ADF, SSIS, Biml and T-SQL

Geeky Things





Fun Things

Active in the PASS community as a speaker, blogger and chronic volunteer.
Likes: coffee, chocolate and cat gifs :)

24HOURS
OF PASS

Pipelines and Packages: Introduction to Azure Data Factory

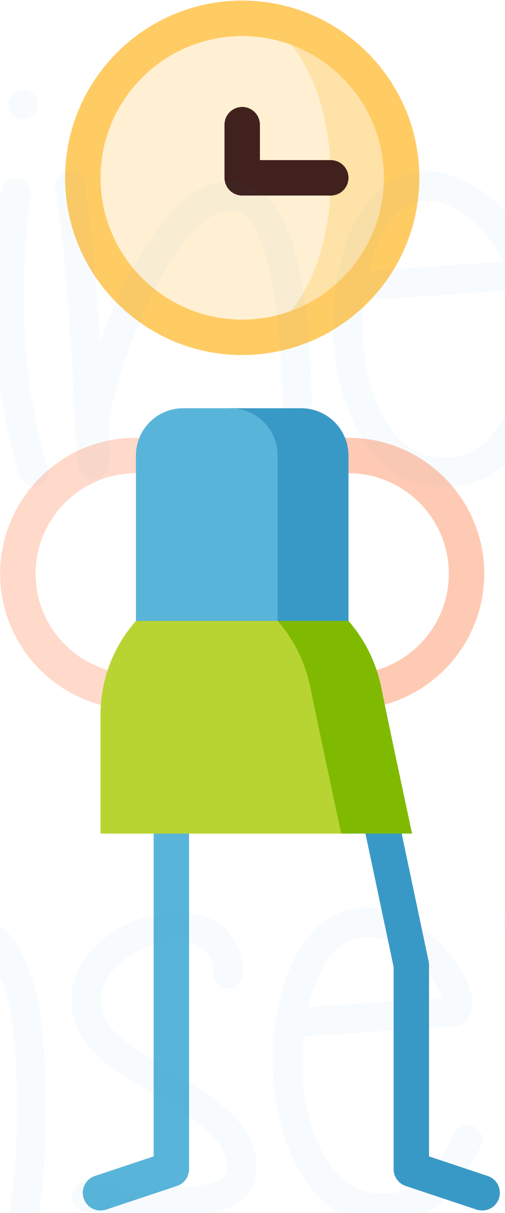
Cathrine Wilhelmsen, Senior BI Consultant, Inmeta
Moderated By: Giuliana Grecco

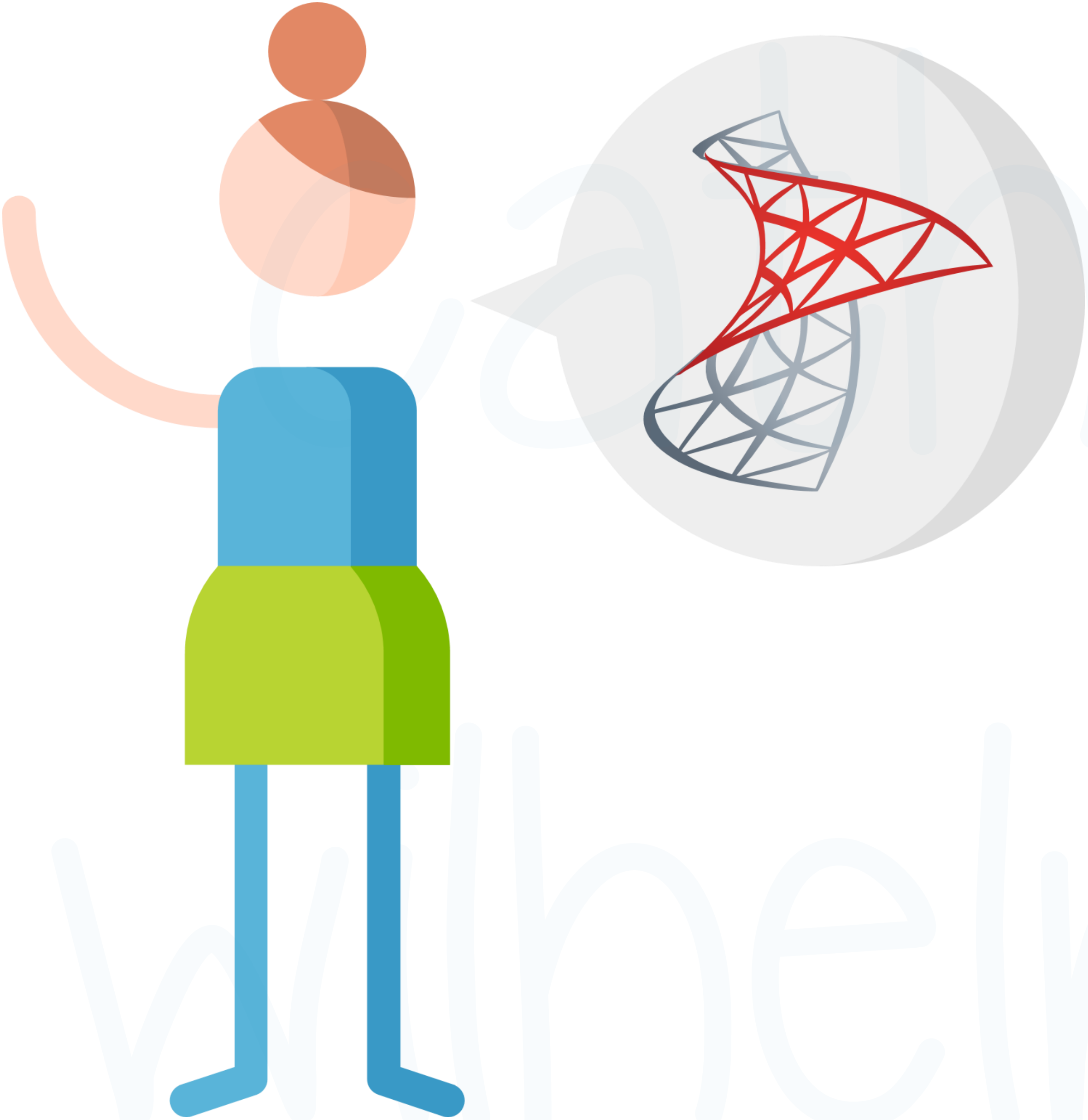
Presenting Sponsors

Microsoft intel

Past Learnings and Future Visions

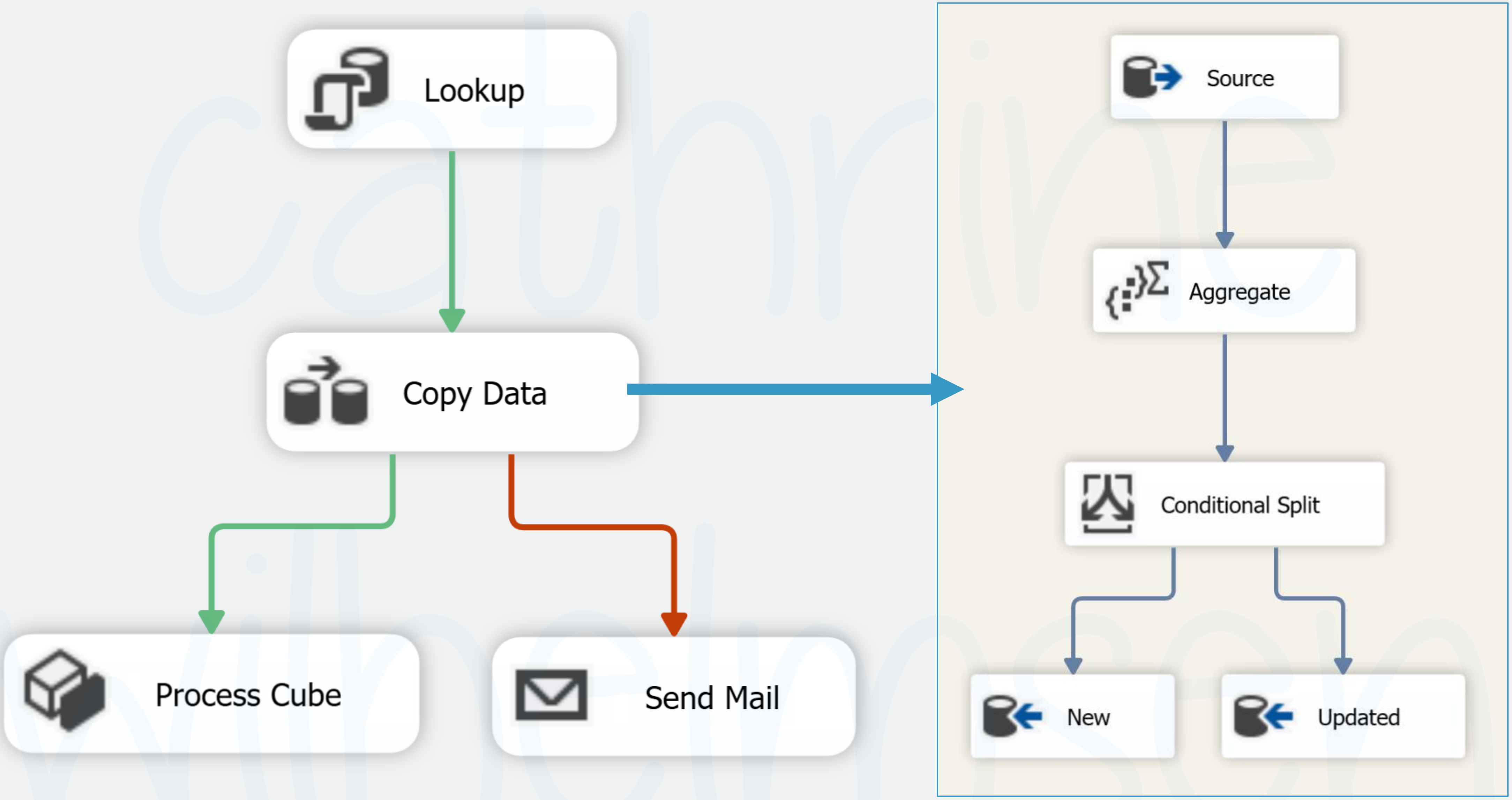
10 years ago...



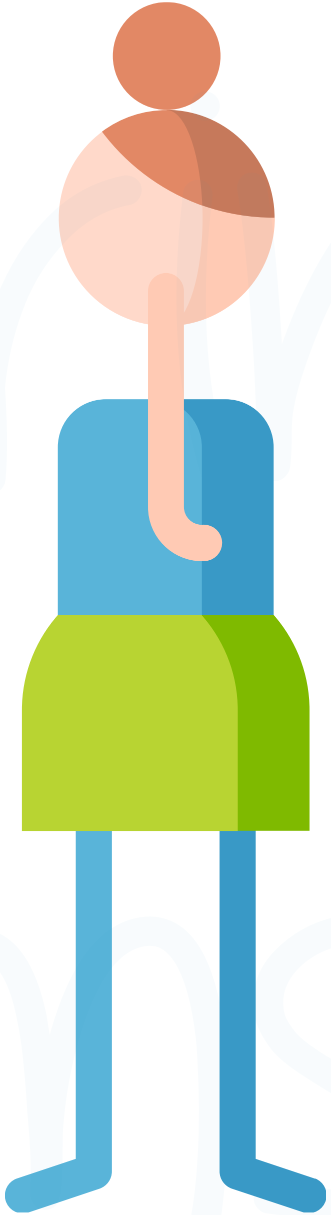


SSIS

SQL Server Integration Services


Lookup
Copy Data
Process Cube
Send Mail
Source
Aggregate
Conditional Split
New
Updated

Then...



ADF v1

Azure Data Factory Version 1

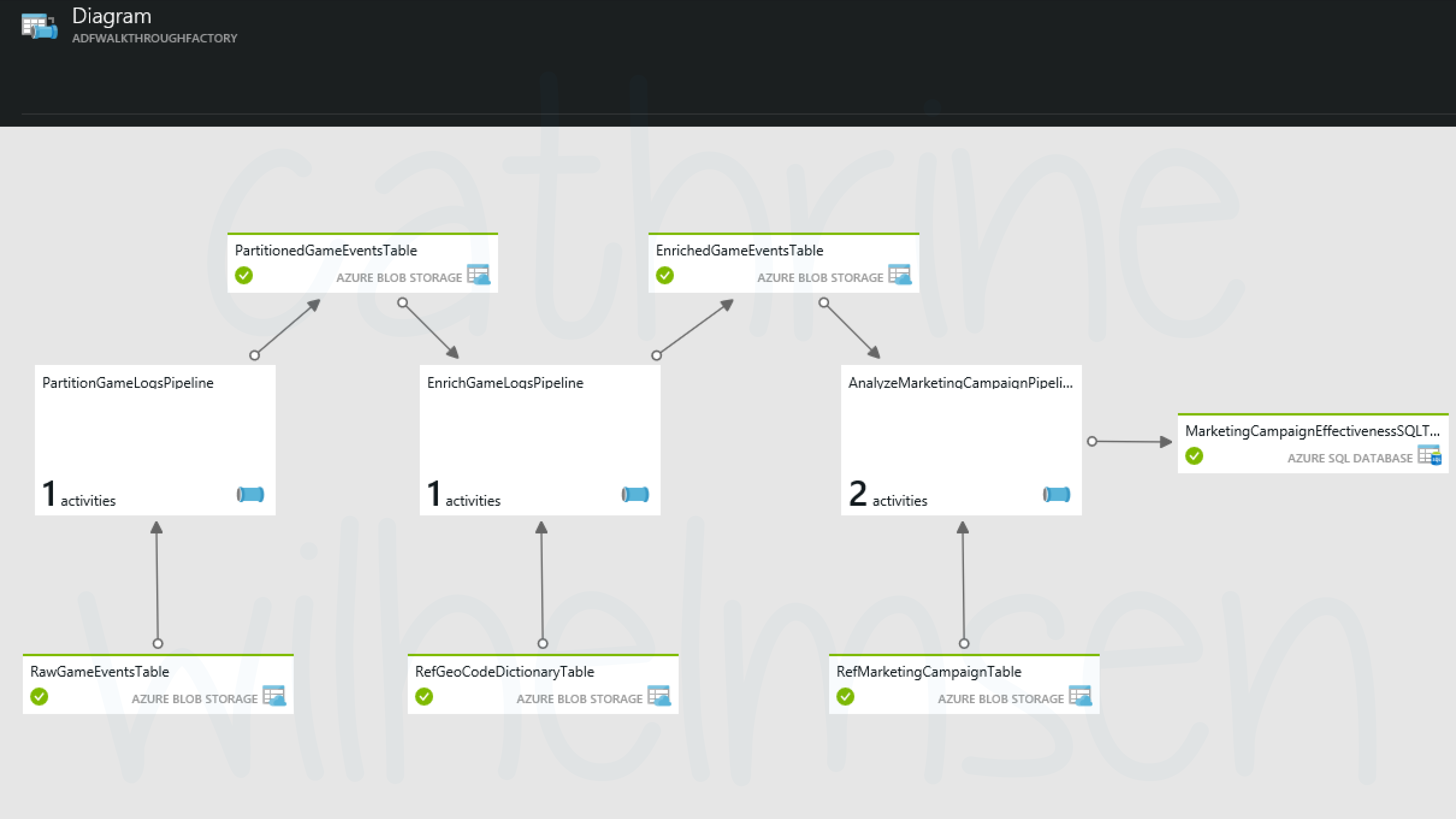
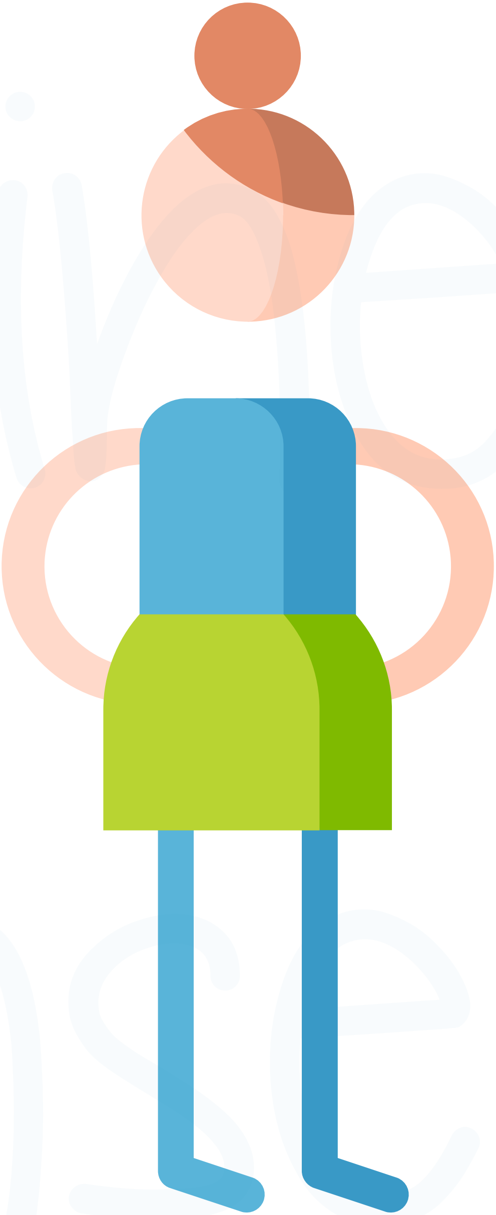
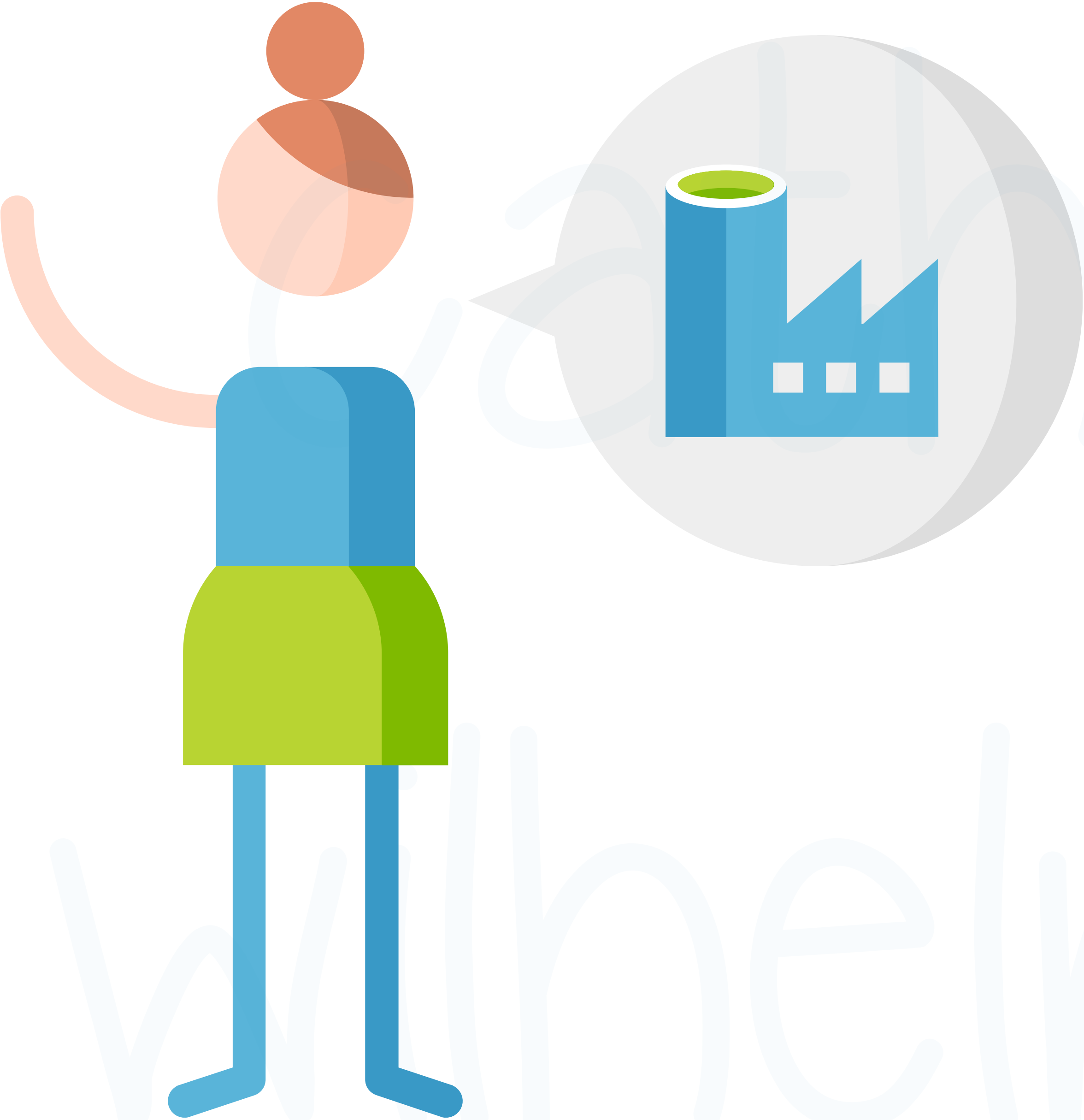


Diagram
ADFWALKTHROUGHFACTORY
PartitionedGameEventsTable
AZURE BLOB STORAGE
EnrichedGameEventsTable
AZURE BLOB STORAGE
PartitionGameLogsPipeline
1 activities
EnrichGameLogsPipeline
1 activities
AnalyzeMarketingCampaignPipeli...
2 activities
MarketingCampaignEffectivenessSQLT...
AZURE SQL DATABASE
RawGameEventsTable
AZURE BLOB STORAGE
RefGeoCodeDictionaryTable
AZURE BLOB STORAGE
RefMarketingCampaignTable
AZURE BLOB STORAGE

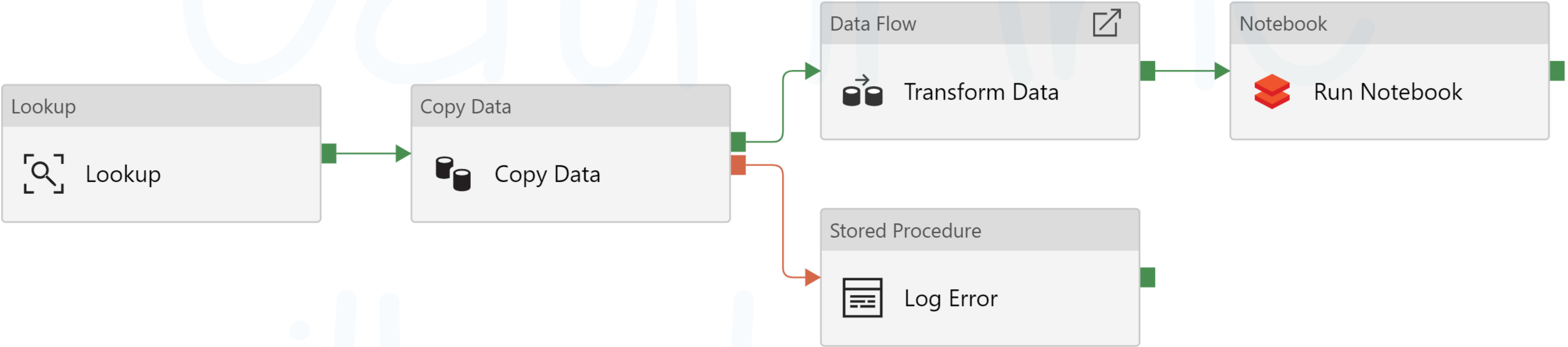
Today…





ADF v2

Azure Data Factory Version 2



Lookup

Lookup

Copy Data

Copy Data

Data Flow

Transform Data

Stored Procedure

Log Error

Notebook

Run Notebook

Stop using SSIS?

Why?

Move to ADF?

How?

What?

Existing solution?

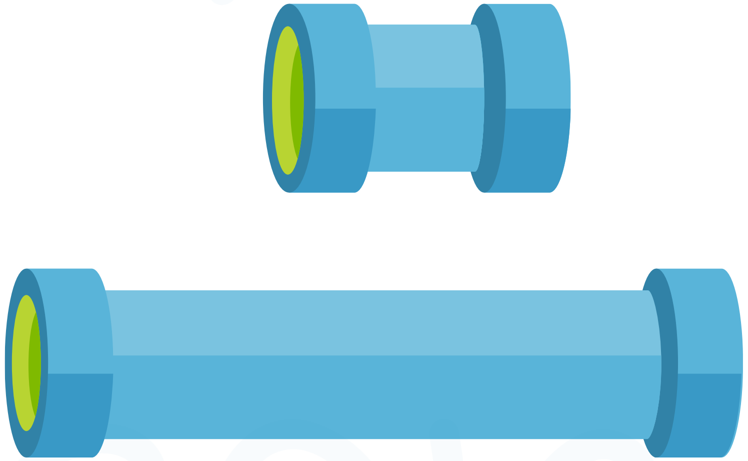
Azure Data Factory

Azure Data Factory

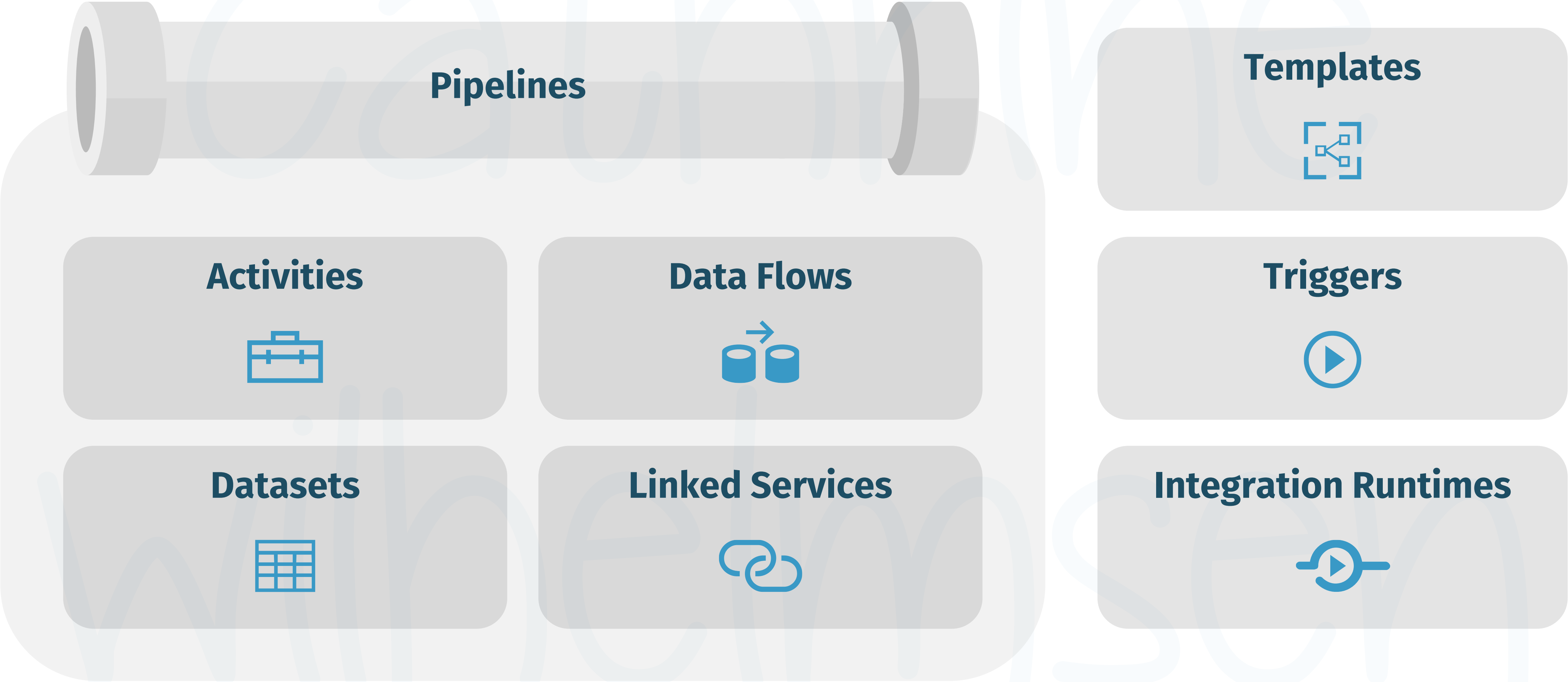
Hybrid data integration service

Complex and scalable pipelines

No-code ETL/ELT workflows



Azure Data Factory



Pipelines

Activities

Data Flows

Datasets

Linked Services

Templates

Triggers

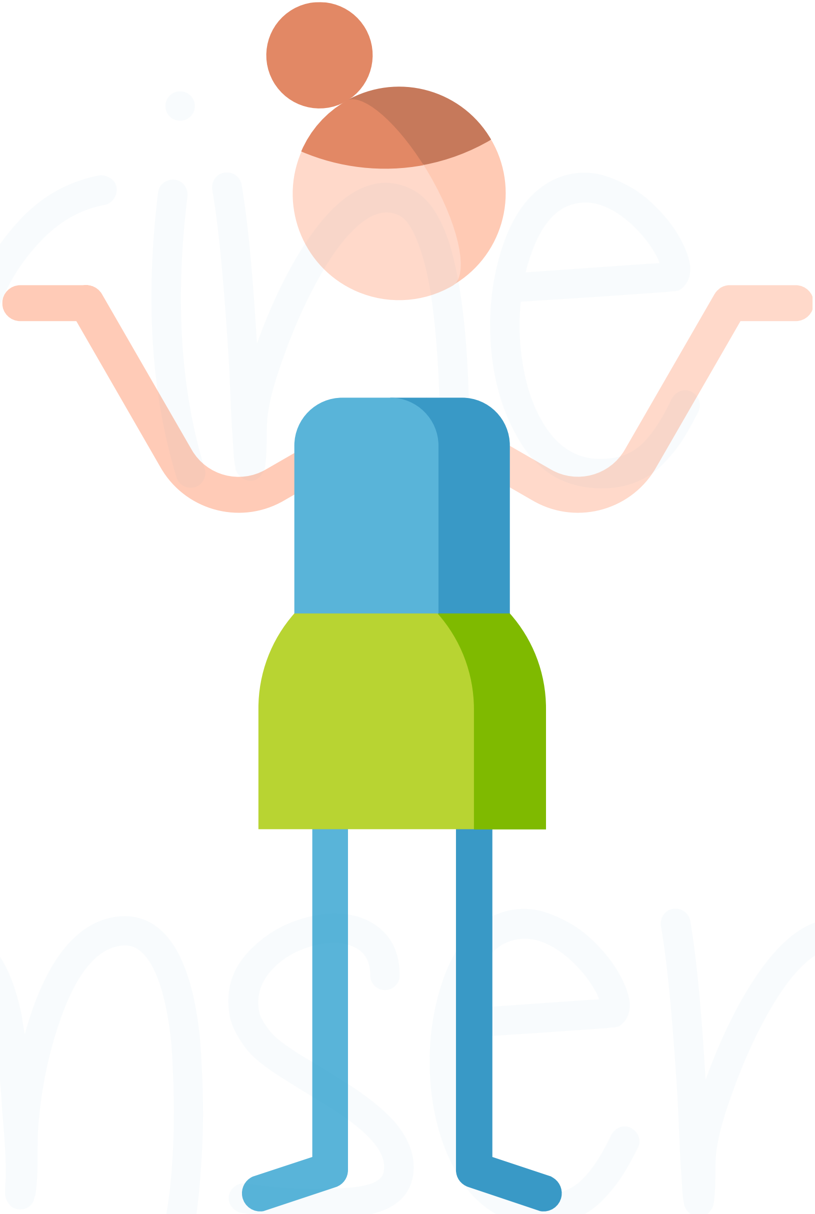
Integration Runtimes

DEMO

Azure Data Factory

Wait…

I already have thousands of SSIS packages!



SSIS
Lift and Shift

Why Lift and Shift SSIS?

Reduce maintenance and costs

Modernize solution while retaining investments

Continue to use familiar tools and processes

How to Lift and Shift SSIS

1. Create Azure SQL Server to host SSISDB
2. Configure SSIS Integration Runtime
3. Deploy SSIS Packages to Azure SQL DB
4. Orchestrate SSIS Packages in Azure Data Factory

SSIS Integration Runtime

Managed cluster of Azure VMs dedicated to SSIS

Customize setup to install third-party components

Join to virtual network for on-premises data access

SSIS Integration Runtime Deep Dive

DEEPER INTEGRATION AND NEW TRANSFORMATION FOR SSIS IN ADF

Sandy Winarko
Principal Program Manager, Azure Data Governance
Cloud + AI, Microsoft

Sandy Winarko
Deeper Integration and New Transformation for SSIS in ADF

https://sqlbits.com/Sessions/Event18/Deeper_Integration_and_New_Transformation_for_SSIS_in_ADF

DEMO

SSIS
Lift and Shift

SSIS Lift and Shift: Lessons Learned

Billed while running (*like all VMs*)

Manage cost by running when necessary

Takes 20-30 minutes to start and stop

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**

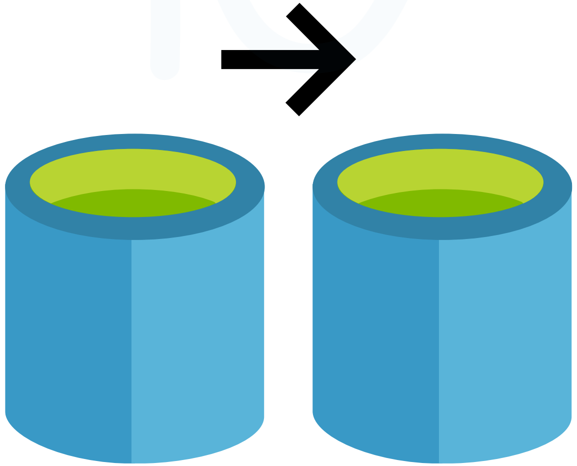
Mapping Data Flows

Mapping Data Flows

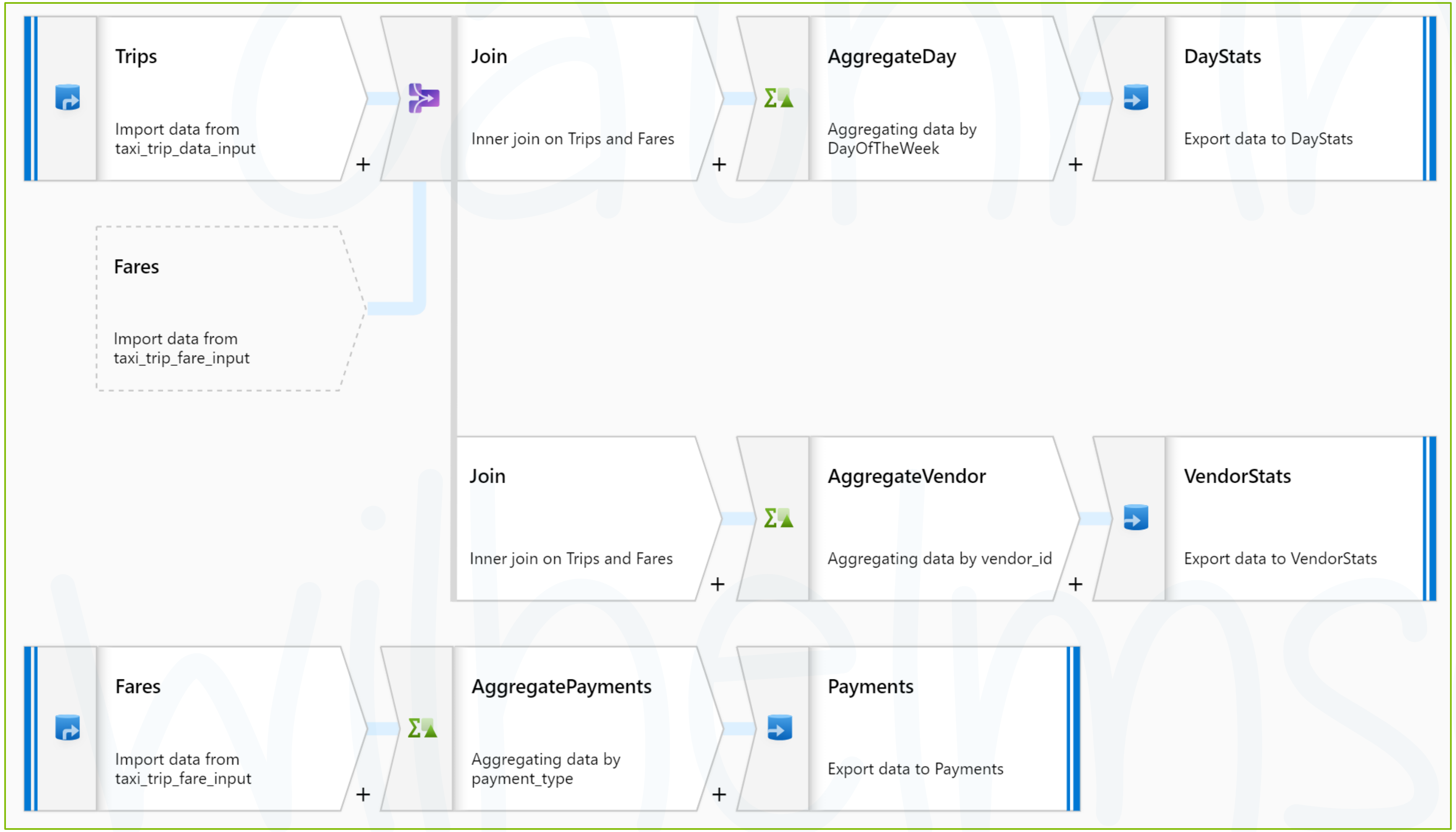
Data transformation at scale

Runs on Azure Databricks

Visual editor, no-code experience



Mapping Data Flows Example



Trips

Import data from taxi_trip_data_input

Join

Inner join on Trips and Fares

AggregateDay

Aggregating data by DayOfTheWeek

DayStats

Export data to DayStats

Fares

Import data from taxi_trip_fare_input

Join

Inner join on Trips and Fares

AggregateVendor

Aggregating data by vendor_id

VendorStats

Export data to VendorStats

Fares

Import data from taxi_trip_fare_input

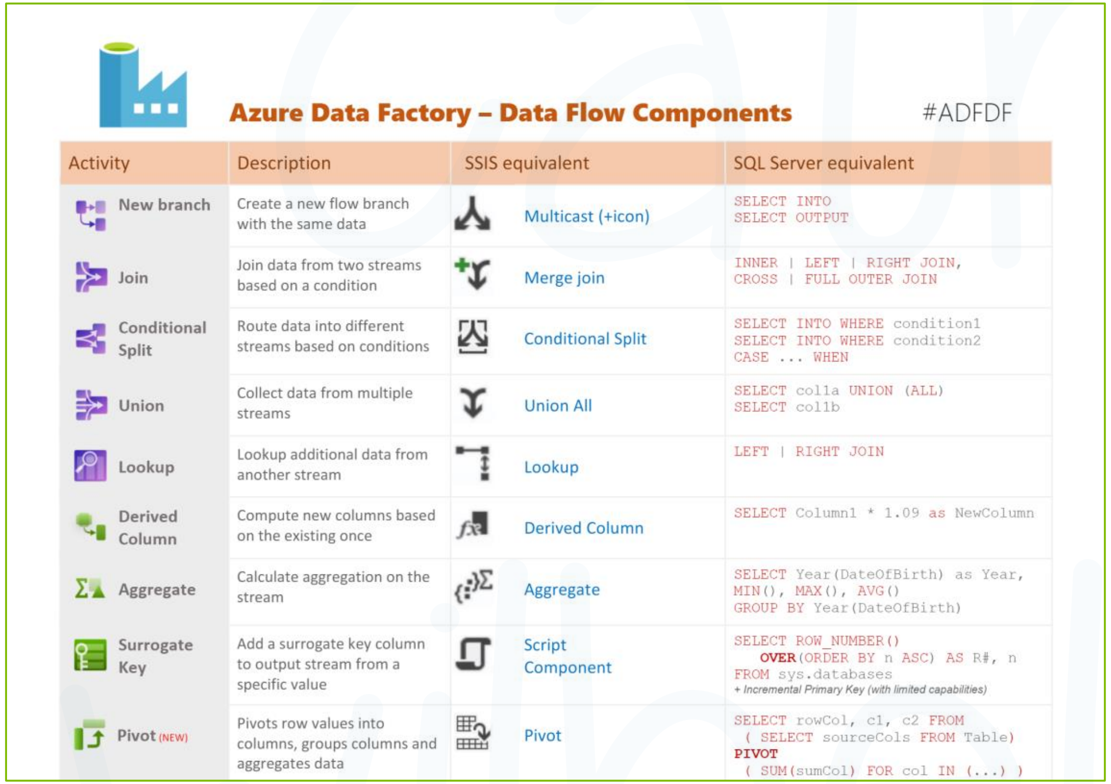
AggregatePayments

Aggregating data by payment_type

Payments

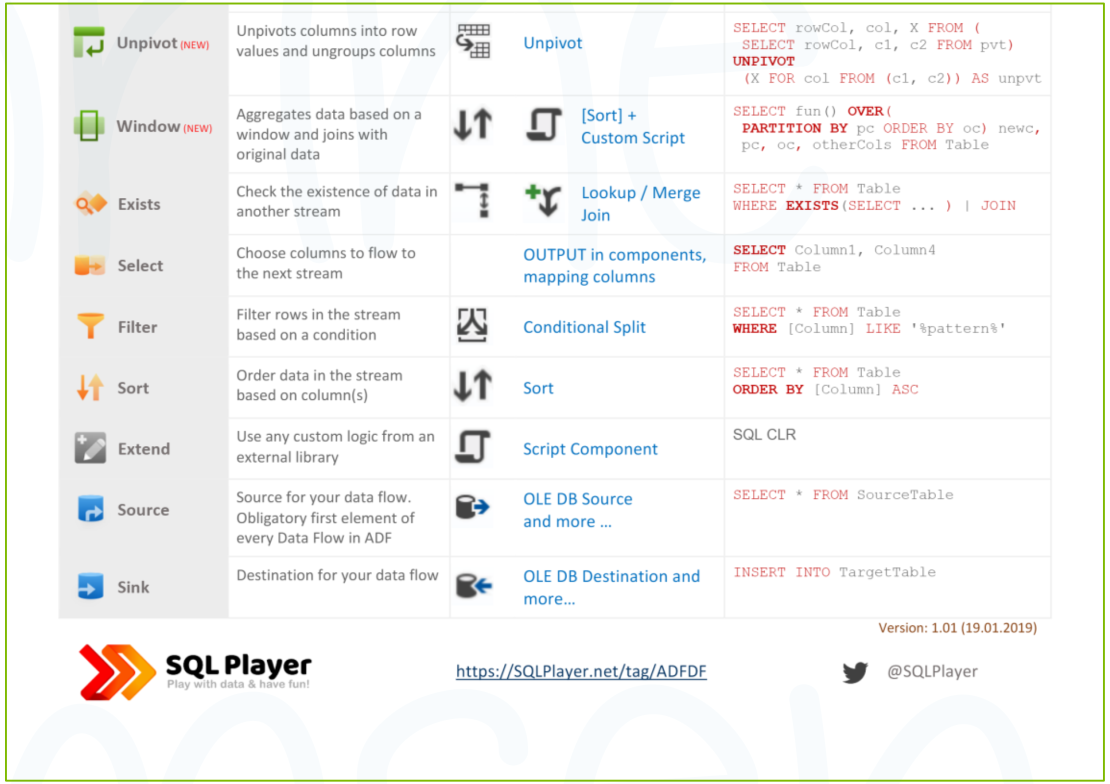
Export data to Payments

Kamil Nowinski's Cheat Sheet



Azure Data Factory – Data Flow Components

#ADFDF



Activity	Description	SSIS equivalent	SQL Server equivalent

New branch	Create a new flow branch with the same data	Multicast (+icon)	SELECT INTO
			SELECT OUTPUT

Join	Join data from two streams based on a condition	Merge join	INNER | LEFT | RIGHT JOIN, CROSS | FULL OUTER JOIN

Conditional Split	Route data into different streams based on conditions	Conditional Split	SELECT INTO WHERE condition1
			SELECT INTO WHERE condition2
			CASE ... WHEN

Union	Collect data from multiple streams	Union All	SELECT col1a UNION (ALL)
			SELECT col1b

Lookup	Lookup additional data from another stream	Lookup	LEFT | RIGHT JOIN
Derived Column	Compute new columns based on the existing once	Derived Column	SELECT Column1 * 1.09 as NewColumn

Aggregate	Calculate aggregation on the stream	Aggregate	SELECT Year(DateOfBirth) as Year,
			MIN(), MAX(), AVG()
			GROUP BY Year(DateOfBirth)

Surrogate Key	Add a surrogate key column to output stream from a specific value	Script Component	SELECT ROW_NUMBER()
			OVER(ORDER BY n ASC) AS R#, n
			FROM sys.databases
			+ Incremental Primary Key (with limited capabilities)

Pivot (NEW)	Pivots row values into columns, groups columns and aggregates data	Pivot	SELECT rowCol, c1, c2 FROM (SELECT sourceCols FROM Table) PIVOT (SUM(sumCol) FOR col IN (...))

Unpivot (NEW)	Unpivots columns into row values and ungroups columns	Unpivot	SELECT rowCol, col, X FROM (
			SELECT rowCol, c1, c2 FROM pvt)
			UNPIVOT
			(X FOR col FROM (c1, c2)) AS unpvt

Window (NEW)	Aggregates data based on a window and joins with original data	[Sort] + Custom Script	SELECT fun() OVER(PARTITION BY pc ORDER BY oc) newc, pc, oc, otherCols FROM Table

Exists	Check the existence of data in another stream	Lookup / Merge Join	SELECT * FROM Table
			WHERE EXISTS(SELECT ...) | JOIN

Select	Choose columns to flow to the next stream	OUTPUT in components, mapping columns	SELECT Column1, Column4
			FROM Table

Filter	Filter rows in the stream based on a condition	Conditional Split	SELECT * FROM Table
			WHERE [Column] LIKE '%pattern%'

Sort	Order data in the stream based on column(s)	Sort	SELECT * FROM Table
			ORDER BY [Column] ASC

Extend	Use any custom logic from an external library	Script Component	SQL CLR
Source	Source for your data flow. Obligatory first element of every Data Flow in ADF	OLE DB Source and more ...	SELECT * FROM SourceTable
Sink	Destination for your data flow	OLE DB Destination and more…	INSERT INTO TargetTable

Version: 1.01 (19.01.2019)

SQL Player
Play with data & have fun!

https://SQLPlayer.net/tag/ADFDF

@SQLPlayer

github.com/SQLPlayer/CheatSheets/blob/master/ADFDF-Cheat-Sheet-sqlplayer.pdf

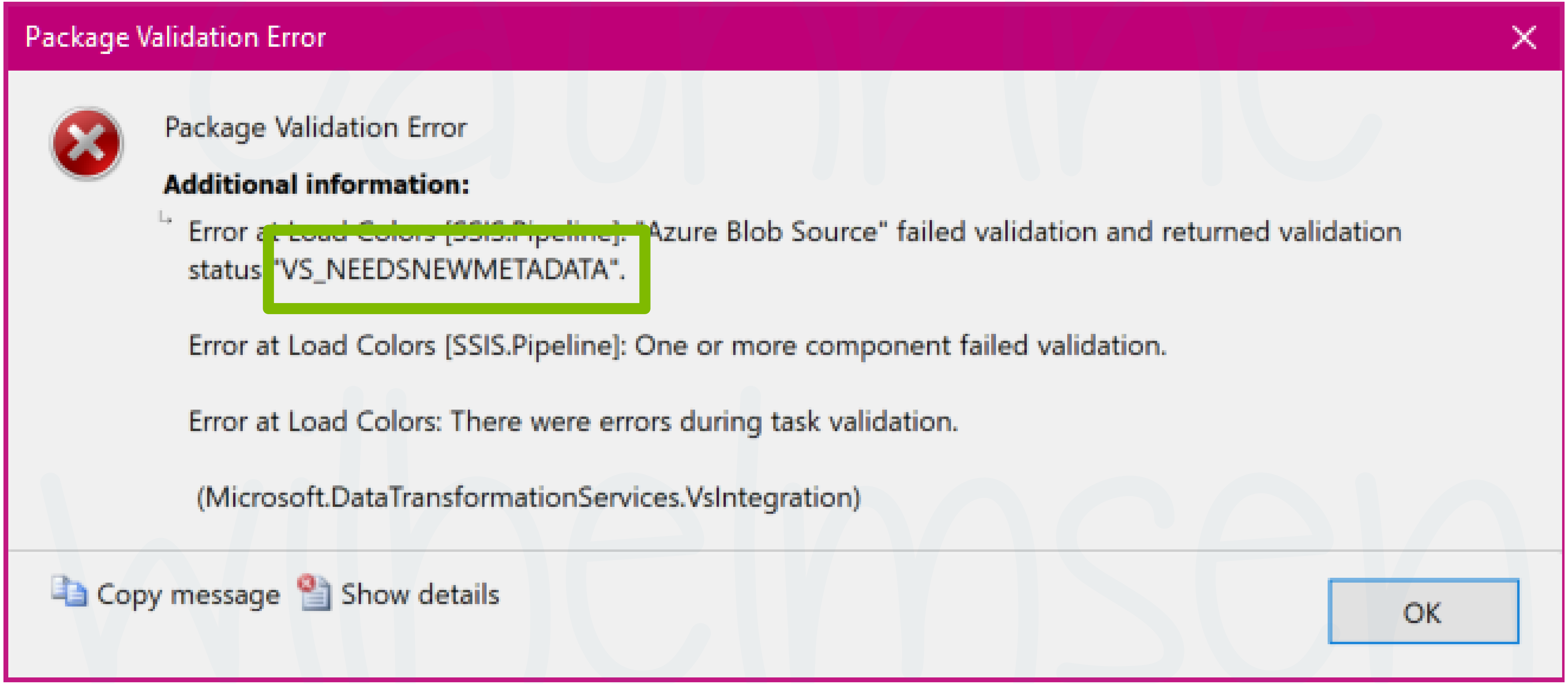
Schema Drift

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



Package Validation Error

Package Validation Error

Additional information:

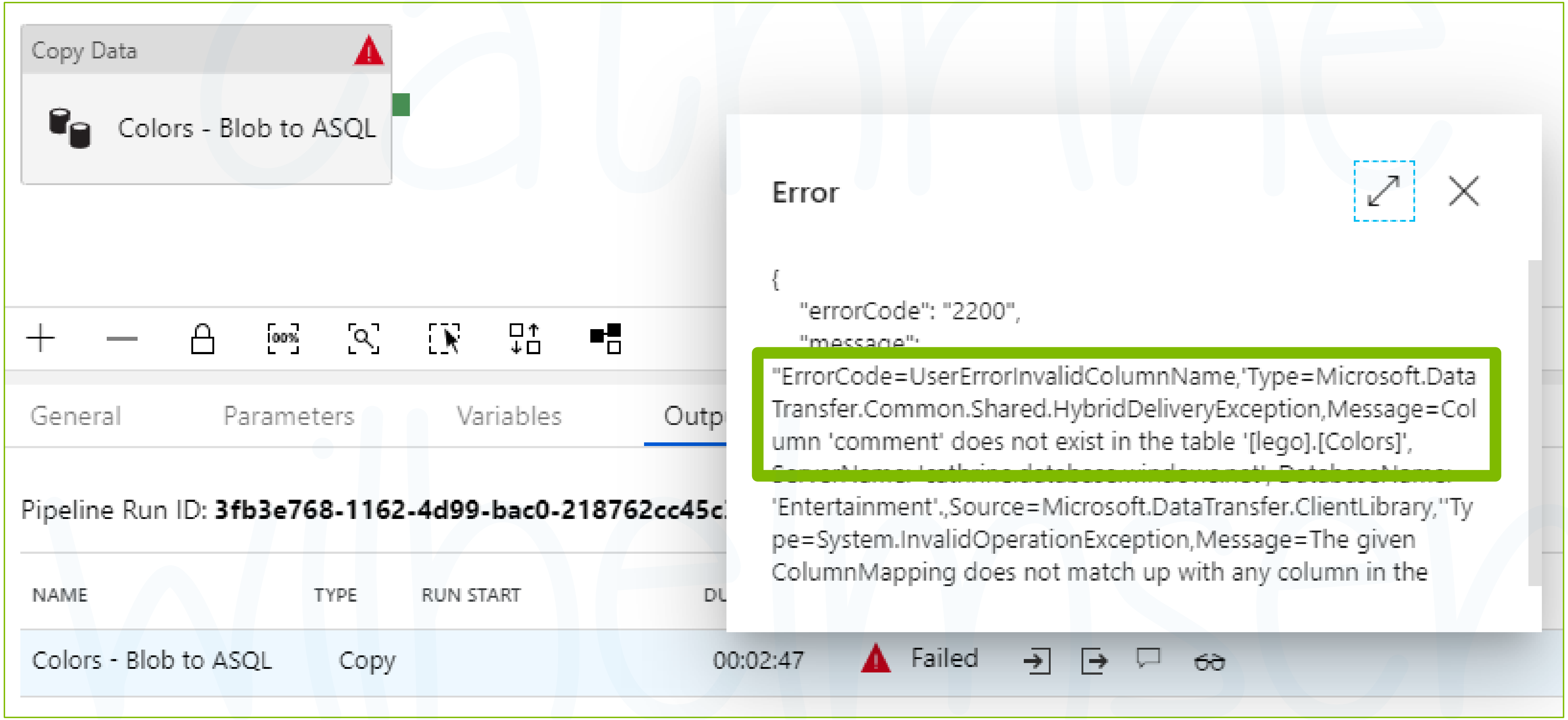
Error at Load Colors [SSIS.Pipeline]: "Azure Blob Source" failed validation and returned validation status "VS_NEEDSNEWMETADATA".

Error at Load Colors [SSIS.Pipeline]: One or more component failed validation.

Error at Load Colors: There were errors during task validation.

(Microsoft.DataTransformationServices.VsIntegration)

Schema Drift in ADF

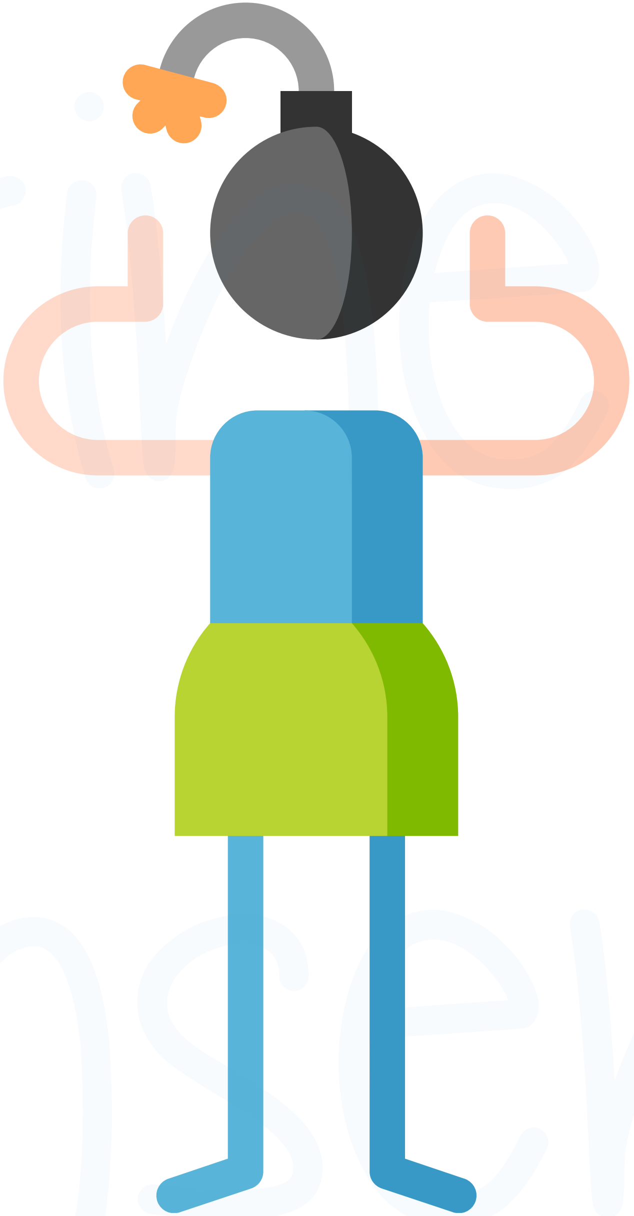


Pipeline Run ID: **3fb3e768-1162-4d99-bac0-218762cc45c**

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

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

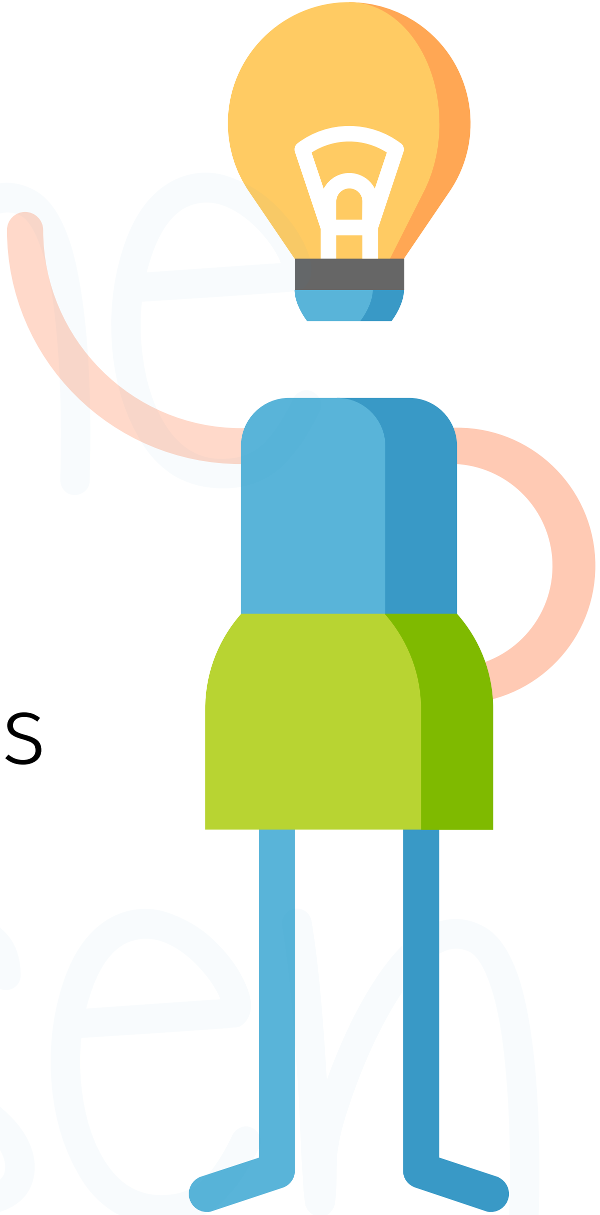
Oh no!



DEMO

Mapping Data Flows

Lessons Learned



In ADF, everything has a price

SSIS best practices != ADF best practices

Learn how to learn and adapt

Good luck!

Thank you :)

Questions?

Microsoft intel

24HOURS
OF PASS

Thank you for attending!

Learn more from Cathrine Wilhelmsen

@cathrinew hi@cathrinew.net

@sqlpass
#sqlpass

@PASScommunity

24HOURS OF PASS

Microsoft intel