Data Community

SQL Day

PLATINUM SPONSOR

TECHNOLOGY INNOVATION DATA KNOWLEDGE tidk

STRATEGIC PARTNER

Microsoft

GOLD SPONSORS

CLOUDS ON MARS

Future Processing

SPYROSOFT

SILVER SPONSOR

dbWatch DATABASE CONTROL

elitmind think bright

VEEAM

BRONZE SPONSOR

Największa księgarnia IT w Polsce

NOVATECH

www.novatech.com.pl

ADF Mapping Data Flow

Tomasz Libera

Tomasz Libera

- Microsoft MVP Data Platform
- Microsoft Certified Trainer
- SQL Server Developer, Trener akademicki
  - WSZiB w Krakowie
- Prowadzi autoryzowane i autorskie szkolenia
  - TSQL | Stored Procedures | Performance Tuning
    - Integration Services | Reporting Services | Power BI
- Data Community
  - Lider Krakowskiej Grupy | Były Członek Zarządu
  - Organizator i prelegent SQLDay, SQLSaturday
- Pasjonat kolarstwa górskiego i maratonów MTB
  - XBOX Bike Team
- tomasz.libera@datacommunity.pl | blog.libera.net.pl

MVP Microsoft® Most Valuable Professional

Agenda

- Azure Data Factory Overview
- Mapping Data Flow
- Transformations
- Demo

Azure Data Factory

- The Azure Data Factory service is a fully managed service for composing data storage, processing, and movement services into streamlined, scalable, and reliable data production pipelines

Azure Data Factory

- Productive
  - Build automated data integration solutions with visual drag-&-drop UI. Move data seamlessly from over **80 sources** without writing code.
- Trusted
  - Data movement using Azure Data Factory has been certified by HIPAA/HITECH, ISO/IEC 27001, ISO/IEC 27018, and CSA STAR.
- Hybrid
  - Build data integration pipelines that span on-premises and cloud. Easily lift your SQL Server Integration Services (SSIS) packages to Azure.
- Scalable
  - Build serverless cloud-based data integration with no infrastructure to manage. Take advantage of elastic capabilities to scale out with your customer growth.

Azure Data Factory

- Visual drag-&-drop UI
  - Maximize productivity by getting pipelines up and running quickly. Use the code-free drag-&-drop interface to build, deploy, monitor, and manage your data integration
- SSIS package execution in Azure
  - Easily execute and schedule your SQL Server Integration Services (SSIS) packages in managed execution environment
- Comprehensive control flow
  - Looping, branching, conditional constructs, on-demand executions, and flexible scheduling
- Multiple Language Support
  - Use the visual interface or write your own code in Python, .NET, or ARM to build pipelines using your existing skills
- Code-free data movement
  - Improve your TCO with 80+ natively supported connectors including Azure data services

New ADF service

Home > New > New data factory

New data factory

* Name
MyAzureDataFactory

* Subscription
Microsoft Azure Sponsorship

* Resource Group
Create new Use existing
MyRG

Version
V2 (Preview)

* Location
East US

Home > MyAzureDataFactory > Settings

MyAzureDataFactory
Data factory

Delete

Essentials

Resource group
MyRG
Location
EastUS
Provisioning state
Succeeded
Getting started
Quick start

Type
Data factory (V2)
Subscription name
Microsoft Azure Sponsorship
Subscription id
14275120-0b96-4eef-98fa-b54e9272fee4

All settings

Quick links

Documentation

Author & Monitor ...

Monitoring

PipelineRuns

100

Settings
MyAzureDataFactory

Filter settings

SUPPORT + TROUBLESHOOTING

Diagnose and solve problems
Activity log
New support request

GENERAL

Properties

GETTING STARTED

Quick start

RESOURCE MANAGEMENT

Tags
Locks

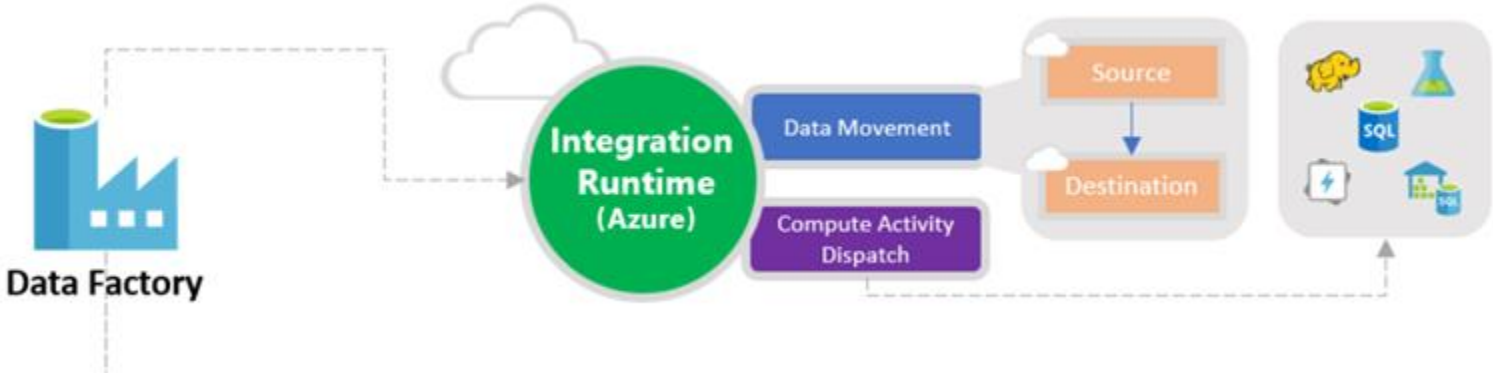
Components

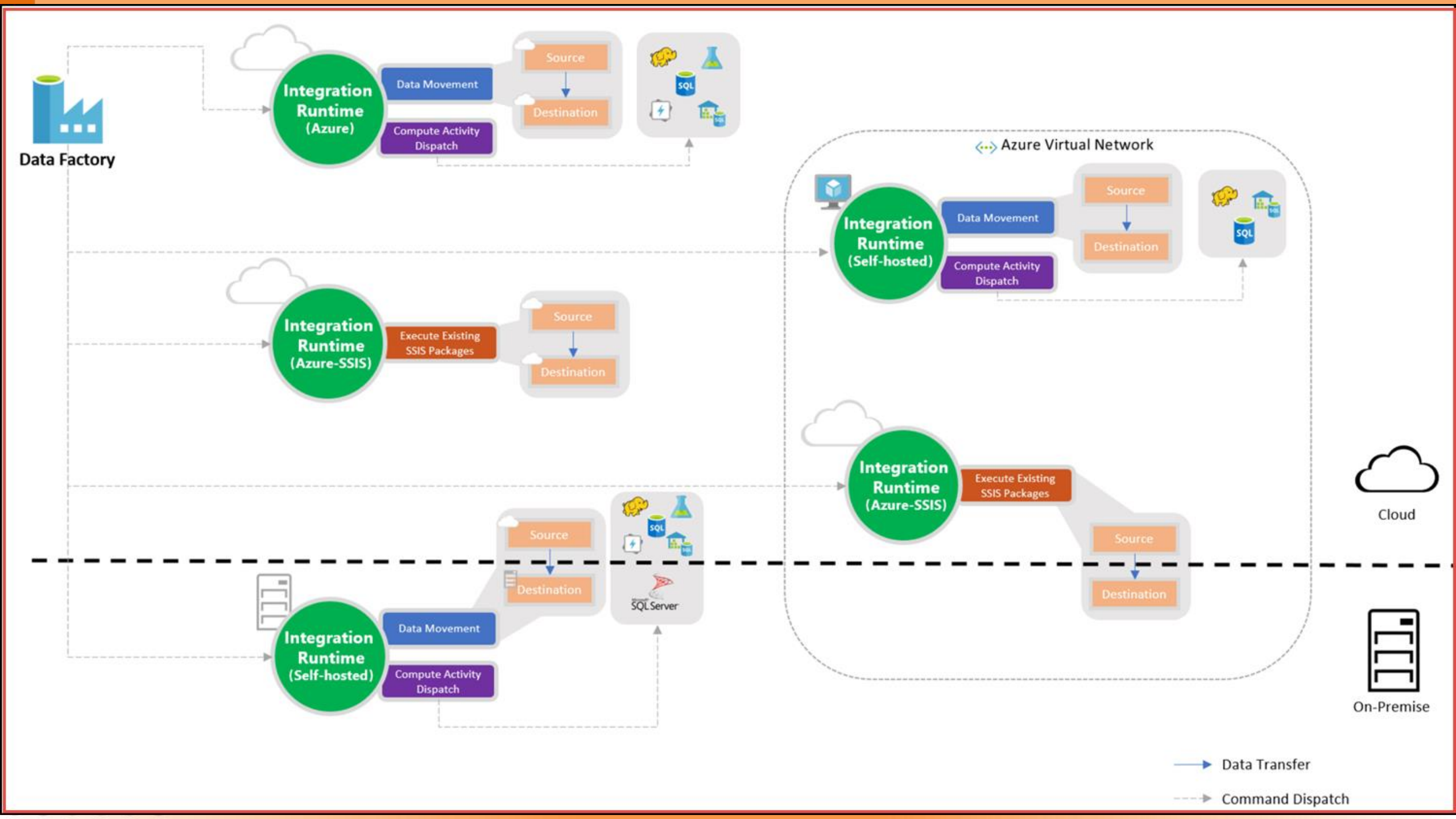
- **Pipeline** – key concept in ADF; workflow of activities
- **Activity** - processing step in a pipeline (copy data, run integration job)
- **Datasets** – data structures
  - point to the data you want to use as inputs or outputs
- **Linked service** – connection string
  - Data store - Azure/ on-premise/ Oracle/ Azure blob storage account
  - Compute resource - execution of an activity – HDInsight Hive/ Pig/ MapReduce/ Streaming
- Triggers - unit of processing
- Parameters

Integration Runtime (IR)

- Data movement
- Activity dispatch - monitor transformation activities
- SSIS package execution: Natively execute SSIS packages

- Types:
- Azure
- Self-hosted
- Azure-SSIS




Data Factory
Integration Runtime (Azure)
Data Movement
Compute Activity Dispatch
Source
Destination
Azure Virtual Network
Integration Runtime (Self-hosted)
Data Movement
Compute Activity Dispatch
Source
Destination
Integration Runtime (Azure-SSIS)
Execute Existing SSIS Packages
Source
Destination
Integration Runtime (Azure-SSIS)
Execute Existing SSIS Packages
Source
Destination
Cloud
Source
Destination
SQL Server
Integration Runtime (Self-hosted)
Data Movement
Compute Activity Dispatch
On-Premise
Data Transfer
Command Dispatch

Azure Data Factory

ADF MAPPING DATA FLOW

Mapping Data Flow

- Transformation capabilities for Azure Data Factory.
  Behind the scene ADF JSON code (generated by drag&drop interface) is converted into Scala programming language and compiled and executed in Azure Databricks which is automatically scale-out as needed.
- Does not require understanding of Spark, Big Data Executions Engines, Clusters, Scala, Python, Java… (Zero-Code)
- Focus on building business logic and data transformations
  - Data cleansing
  - Aggregation
  - Data conversion
  - Data preparation
  - Data exploration

Data Flows vs SSIS

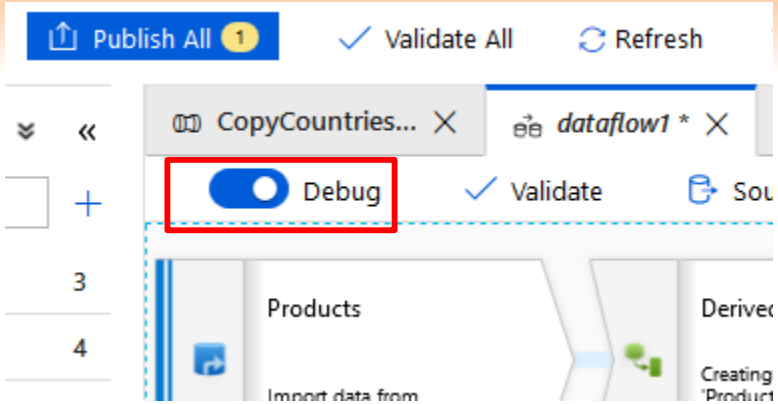
SQL Player
Play with data & have fun!

Operation / 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	*Subselect, function,* `LEFT/RIGHT JOIN`
Derived Column	Compute new columns based on the existing once	Derived Column	`SELECT Column1 * 1.09 as NewColumn`
Exists		Lo[illegible]	
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`

Debug Mode



- Interactively build data flow with a running Azure Databricks interactive cluster
- Session will close once you turn debug off in ADF.
- Hourly charges incurred by Azure Databricks during the time that the debug session turned on

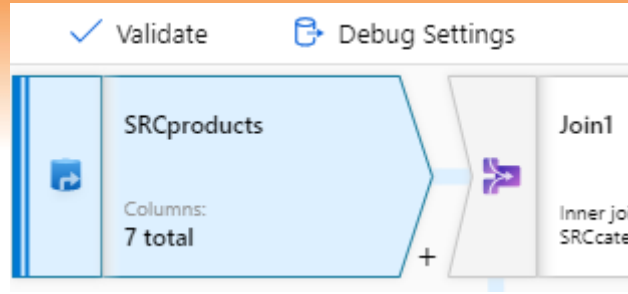
- https://github.com/kromerm/adfdataflowdocs/blob/master/Concepts/adf-data-flow-debug-mode.md

Expression Builder

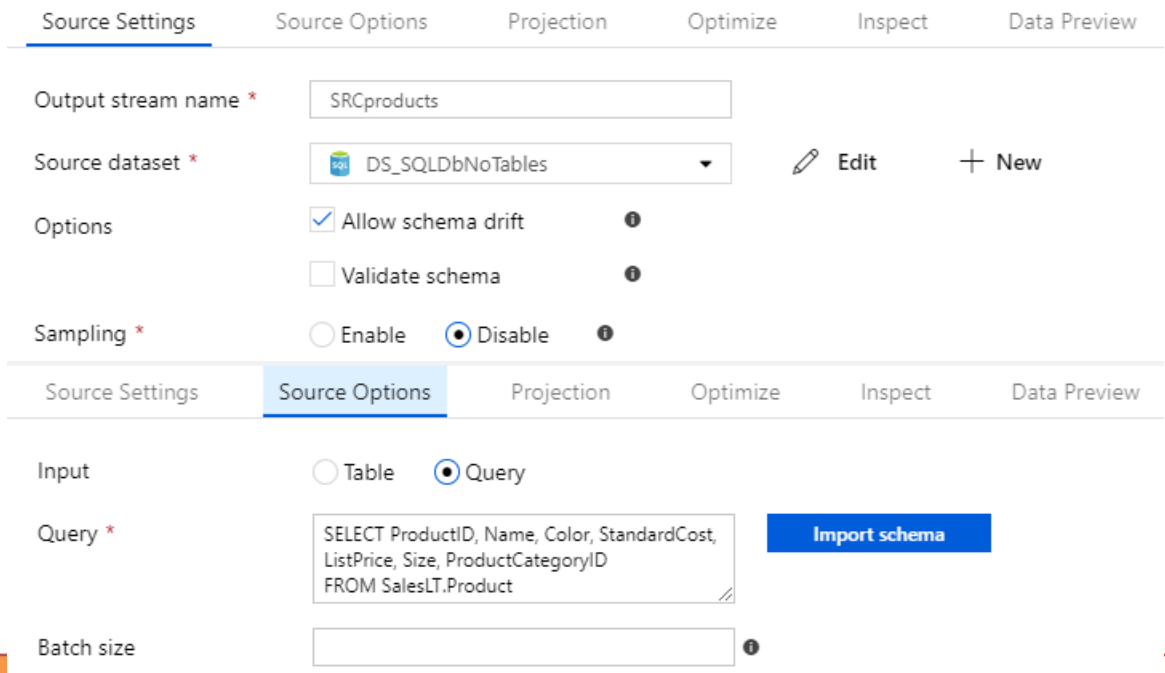
- Some transformations use expressions which could be created in Expression Builder.
- Expressions use columns, fields, variables, parameters and functions.
- Auto-complete feature reads from the entire Azure Data Factory Data Flow object model with syntax checking and highlighting.
- Using Debug mode, live in-progress preview data results and real-time live debugging is enabled.

Transformations

Source



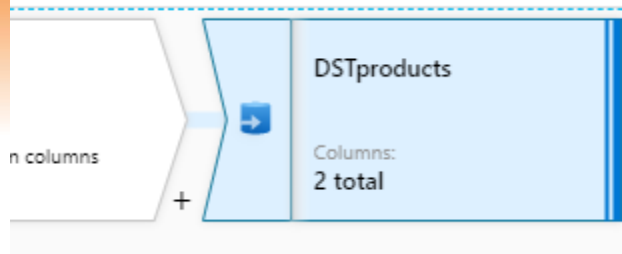
- Data flow can include more than one source.
- Every data flow requires at least one source transformation.
- You can join those sources together with a join transformation or a union transformation.
- Dataset defines the shape and location of the data you want to write to or read from.
- You can use wildcards and file lists in your source to work with more than one file at a time.

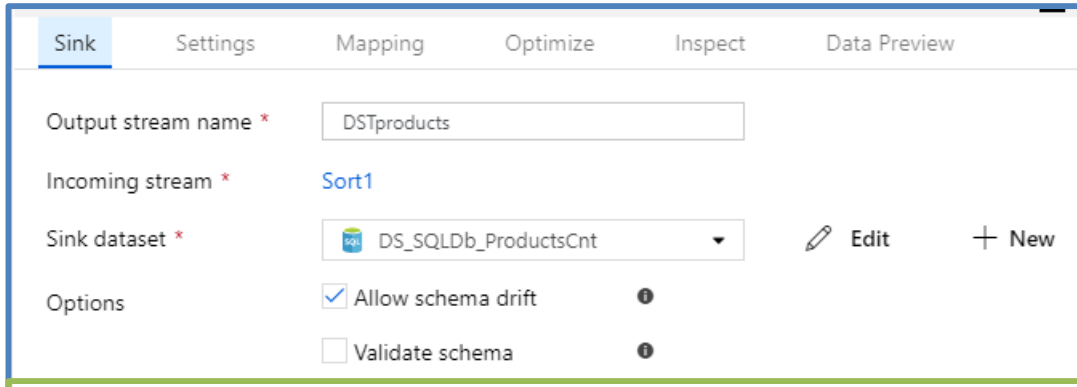


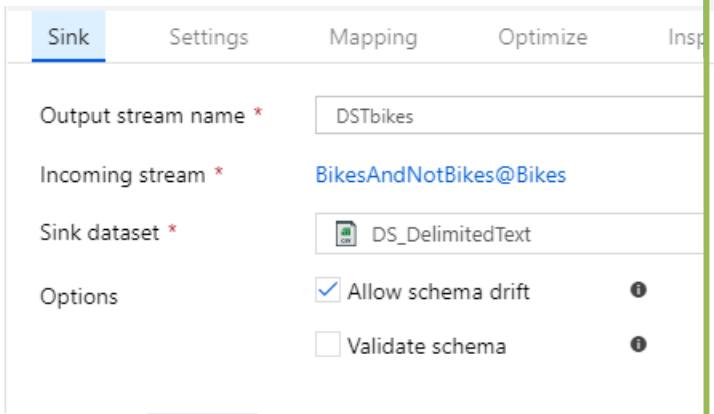
Transformations

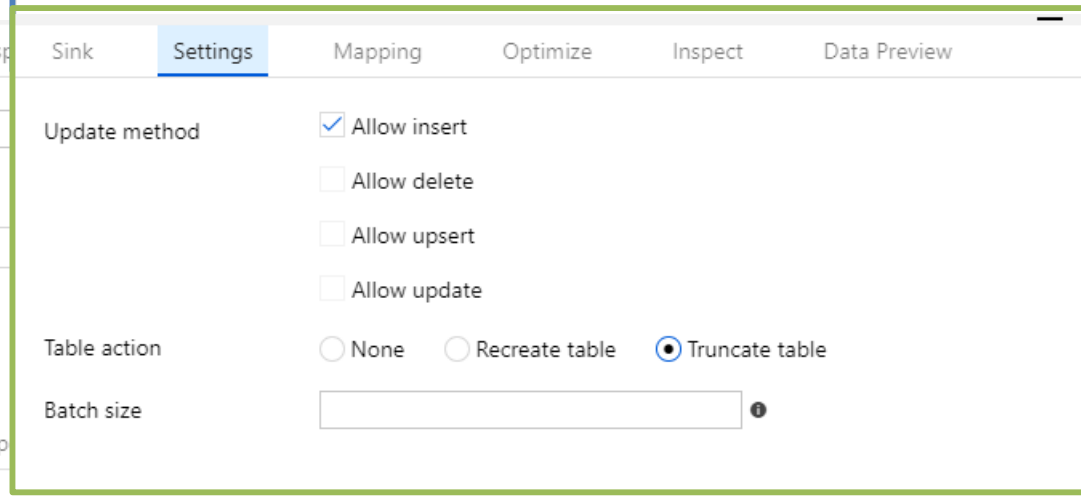
Sink

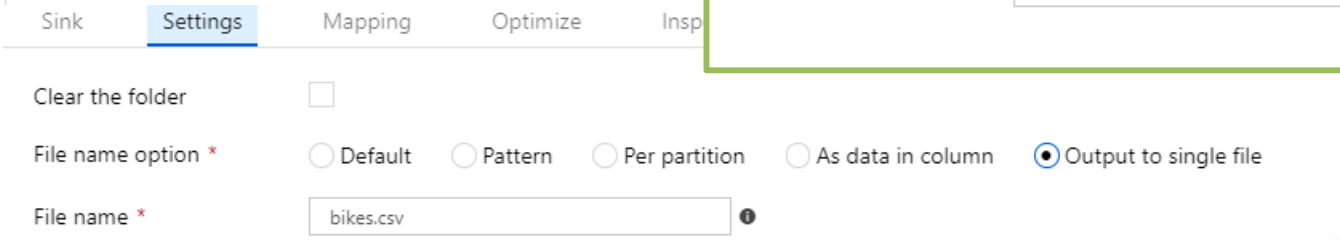
- Data flow can include more than one sink (destination).
- File name options
  - Default
  - Pattern
  - Per partition
  - As data in column
- Database options
  - Update method
  - Recreate table
  - Truncate table
  - Batch size
  - Enable staging







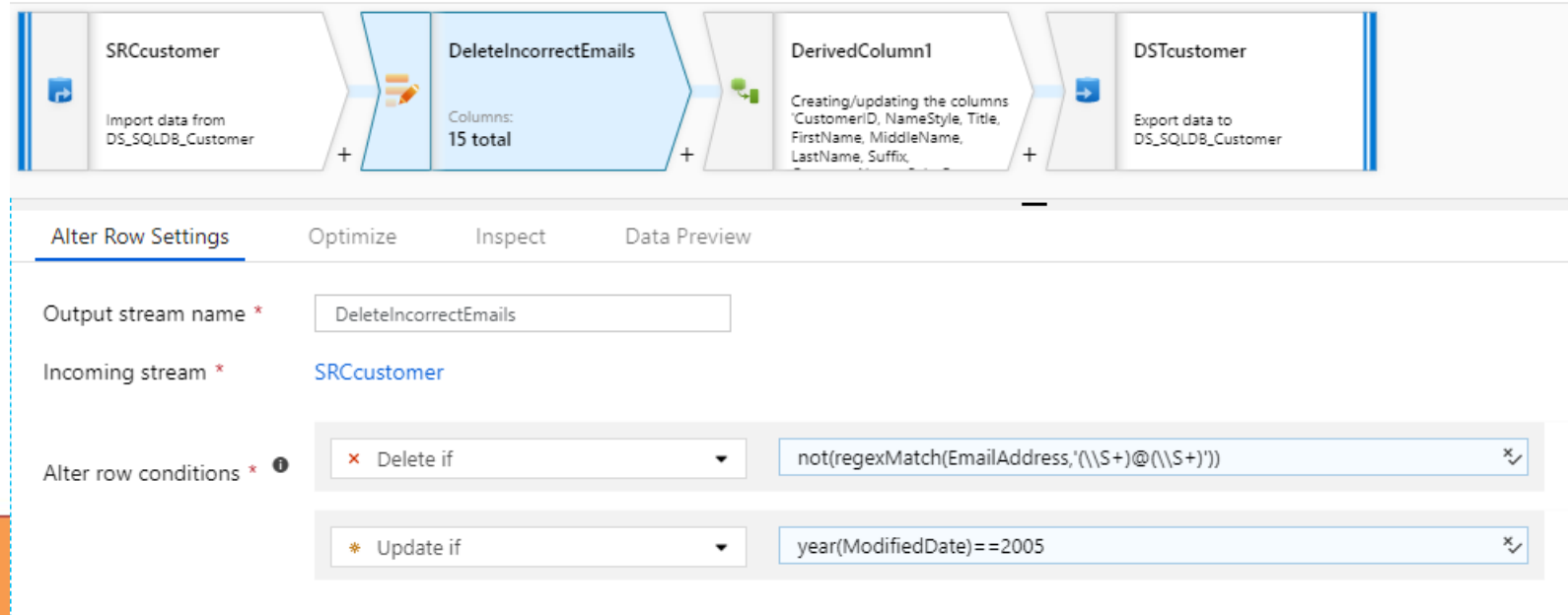




Transformations

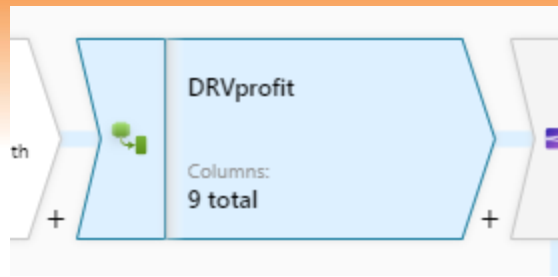
Alter Row

- To set insert, delete, update, and upsert policies on rows.
- Based on expressions (conditions) rows will be inserted, updated, deleted, or upsert
- Alter Row can produce both DDL & DML actions against database.

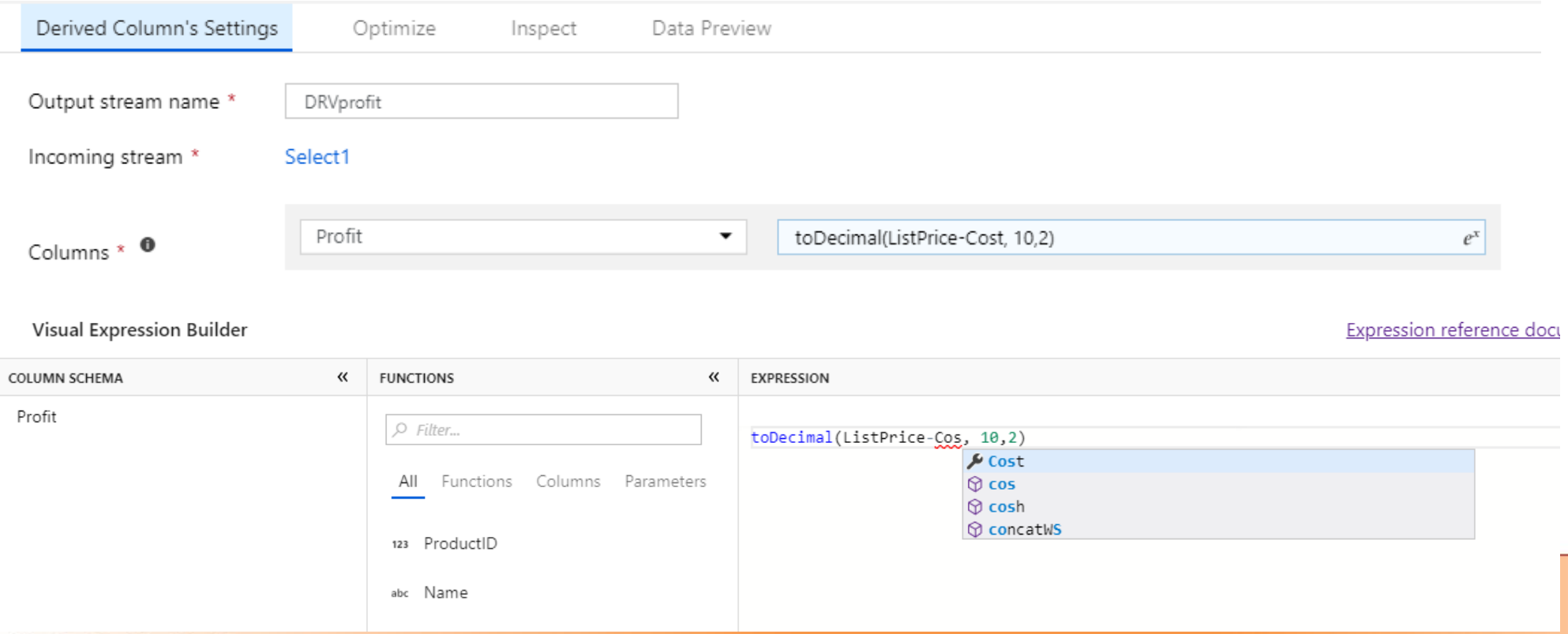


Transformations

Derived Column

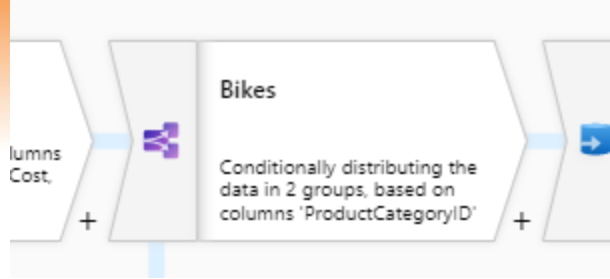


- Generate new columns or to modify existing fields.
- Expression Builder window to build the expression for the derived columns using expression functions.

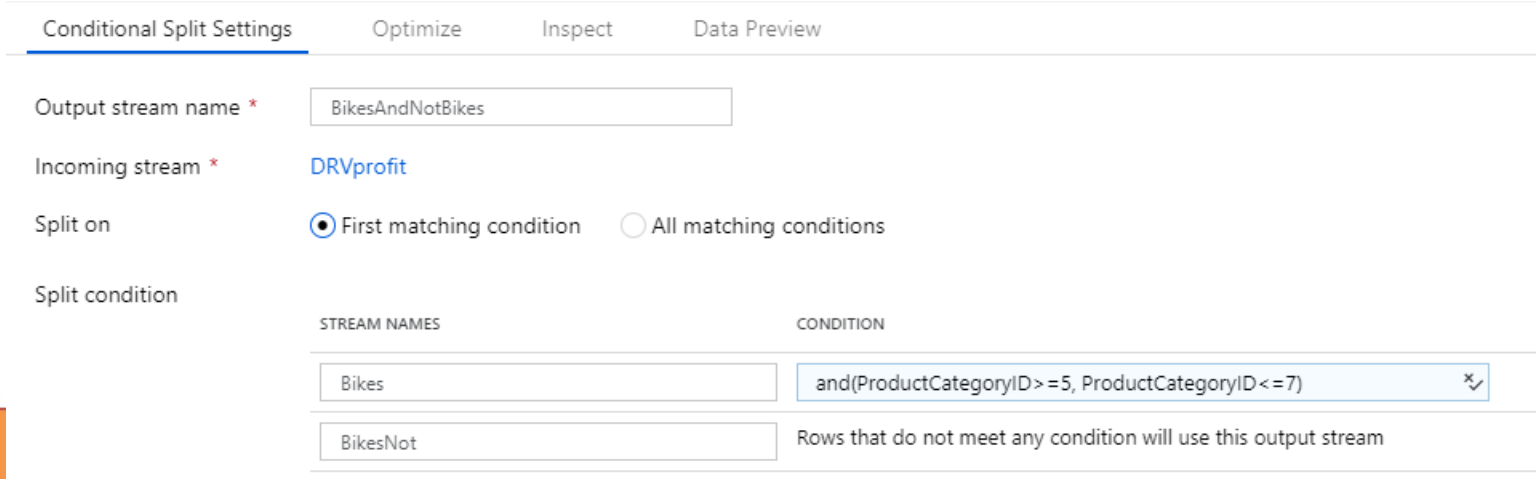


Transformations

Conditional Split

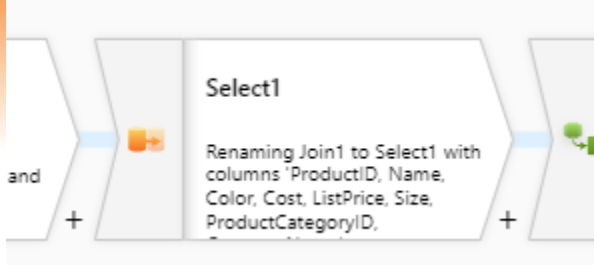


- Generate new columns or to modify existing fields.
- Expression Builder window to build the expression for the derived columns using expression functions.

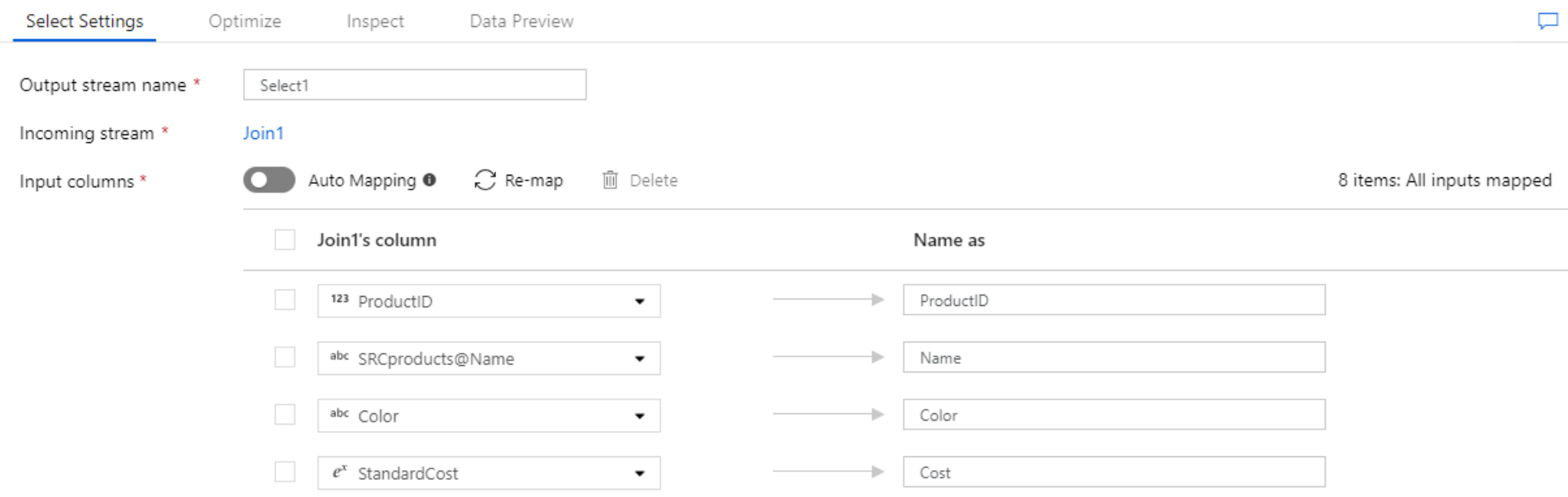


Transformations

Select



- Column selectivity (reducing number of columns)
- Alias columns and stream names

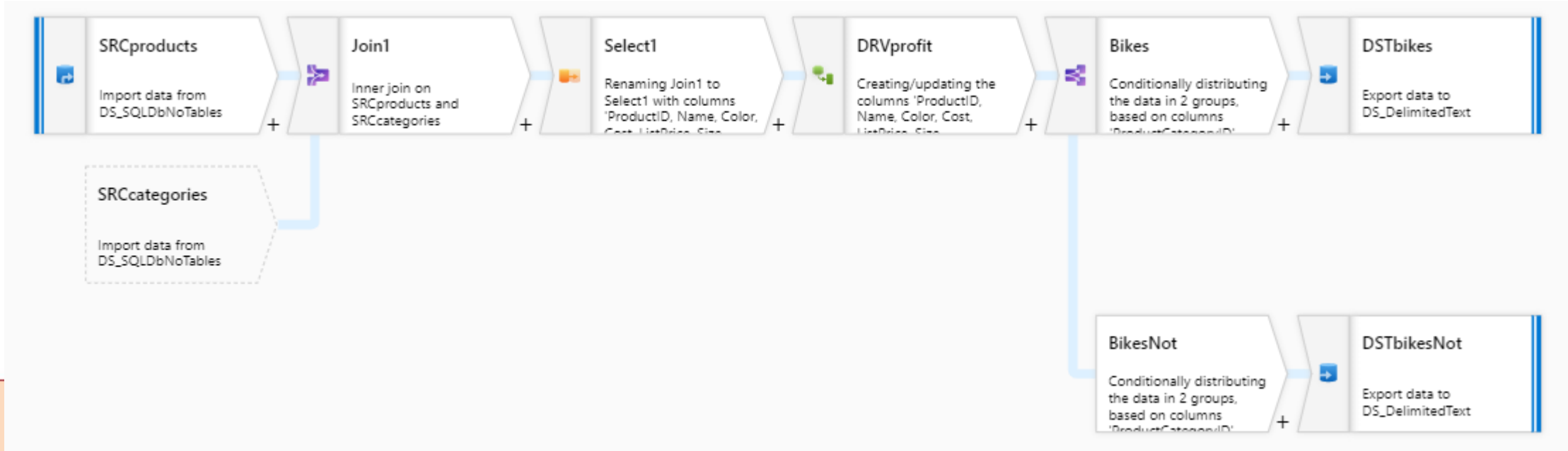


DEMO

A. Load products from table, join with category names and split into bikes and other products.

B. Load products.csv, aggregate – grouping by category name, sort and save results into new table.

C. Sync data in table using Alter Row (delete incorrect email).



Data Community

SQL Day

PLATINUM SPONSOR

TECHNOLOGY INNOVATION DATA KNOWLEDGE tidk

STRATEGIC PARTNER

Microsoft

GOLD SPONSORS

CLOUDS ON MARS

Future Processing

SPYROSOFT

SILVER SPONSOR

dbWatch DATABASE CONTROL

elitmind think bright

VEEAM

BRONZE SPONSOR

Największa księgarnia IT w Polsce

NOVATECH

www.novatech.com.pl

DZIĘKUJĘ ZA UWAGĘ

tomasz.libera@datacommunity.pl

@tomasz_libera

bit.ly/sqlday2019_adf