Azure Data Factory vs SSIS

May the best tool win!

Regis Baccaro

@regbac http://theblobfarm.wordpress.com

GOLD SPONSORS

Atos

Profisee®

Zerto

SILVER SPONSORS

PROMISE
GRUPA APN PROMISE

ELITMIND

ITMAGINATION
Imagine it. Done.

THEOBALD
SOFTWARE

BRONZE SPONSORS

it KONTRAKT®

JCommerce

IT
Krajowy Rejestr Długów

KSIĄŻKI IT Z CAŁEGO ŚWIATA
NOVATECH
www.novatech.com.pl

PEPCO

dbWatch
DATABASE CONTROL

querona

Data Scientist as a Service

UBS

VOLVO
Volvo Group IT

STRATEGIC PARTNER

Microsoft

About.me : Regis Baccaro

Consultant

Developer

Speaker

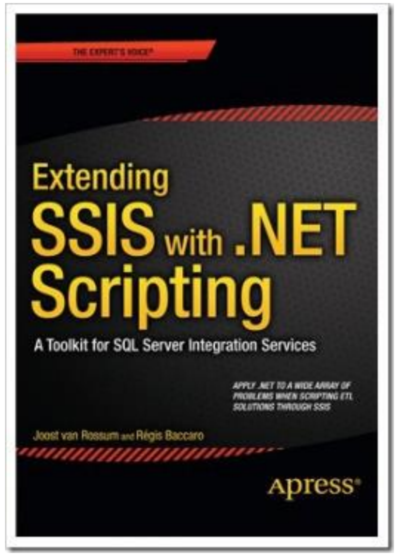
Author

Data Platform MVP

Farmer

SQL Nexus lead

MCT



THE EXPERT'S VOICE®

Extending SSIS with .NET Scripting

A Toolkit for SQL Server Integration Services

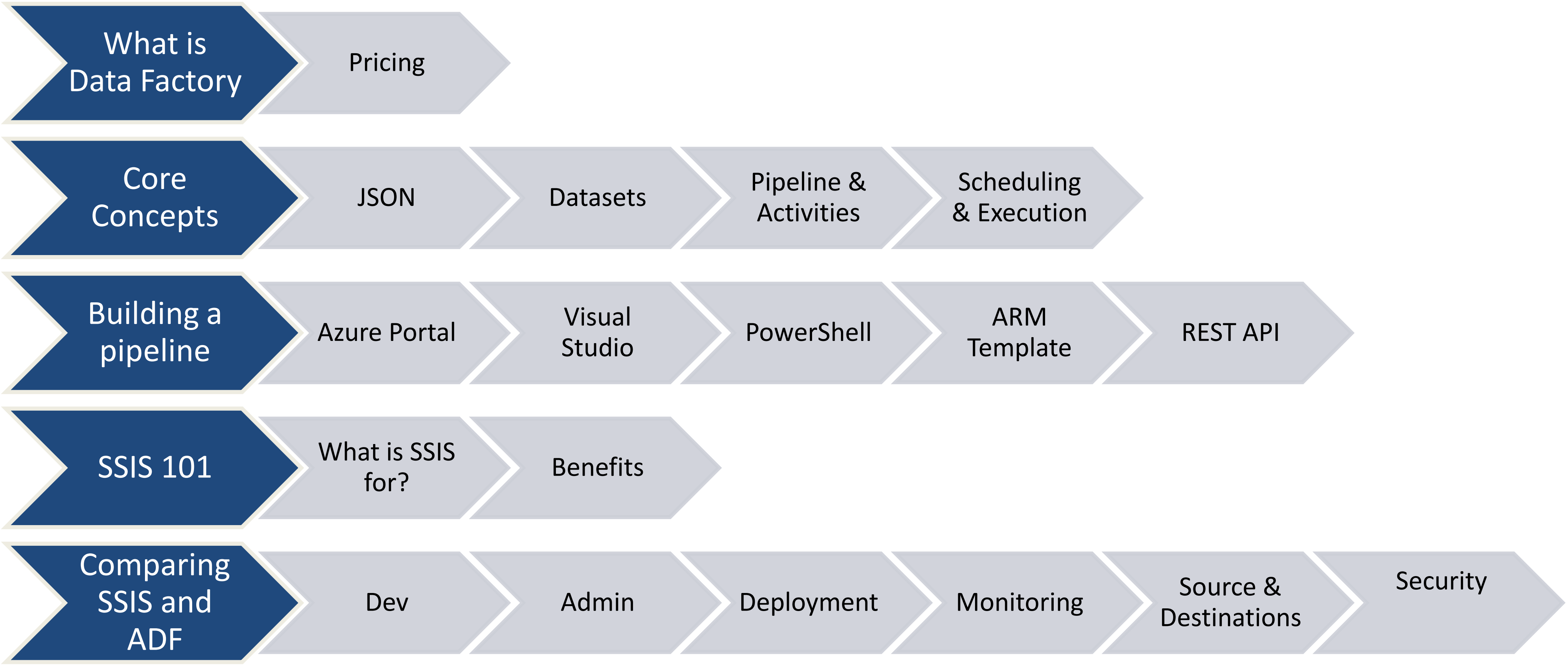
APPLY .NET TO A WIDE ARRAY OF PROBLEMS WHEN SCRIPTING ETL SOLUTIONS THROUGH SSIS

Joost van Rossum and Régis Baccaro

Apress®

The contestants

Agenda



- What is Data Factory
  - Pricing
- Core Concepts
  - JSON
  - Datasets
  - Pipeline & Activities
  - Scheduling & Execution
- Building a pipeline
  - Azure Portal
  - Visual Studio
  - PowerShell
  - ARM Template
  - REST API
- SSIS 101
  - What is SSIS for?
  - Benefits
- Comparing SSIS and ADF
  - Dev
  - Admin
  - Deployment
  - Monitoring
  - Source & Destinations
  - Security

What is Data Factory

Data Factory is a fully managed cloud-based data integration service that orchestrates and automates the movement and transformation of data

Biggest mistake / pain points about Data Factory

Positioning it as SSIS for the Cloud

Adhoc execution

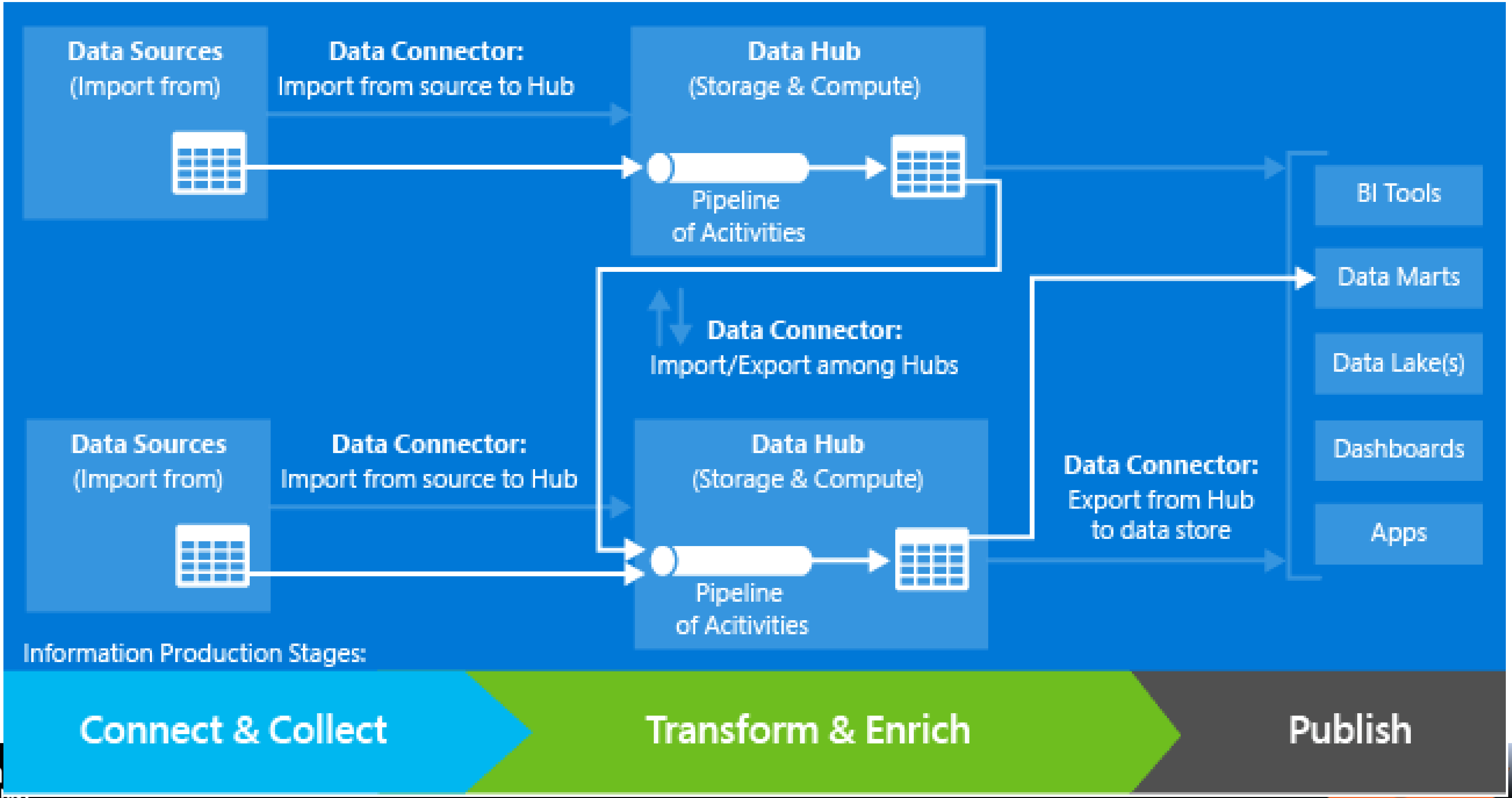
Datatype management

Datasets without time slices

Debugging

Continuous integration

Azure Data Factory Architecture



Pricing

No upfront cost

No termination fees

Pay as you g(r)o(w)

Pay for movement and Data usage

A Dataset is a logical description of the data

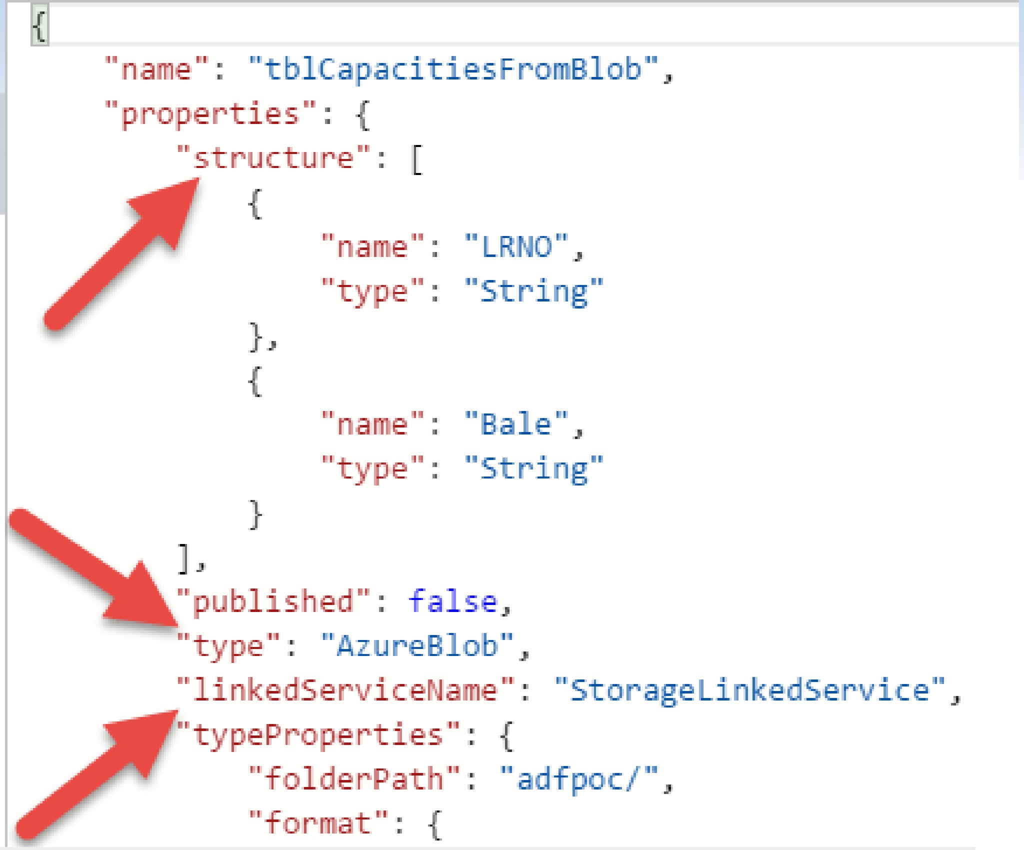
The mechanism (address, protocol, authentication scheme) to access the data is defined in the Linked Service and referenced in the dataset definition.

Logical

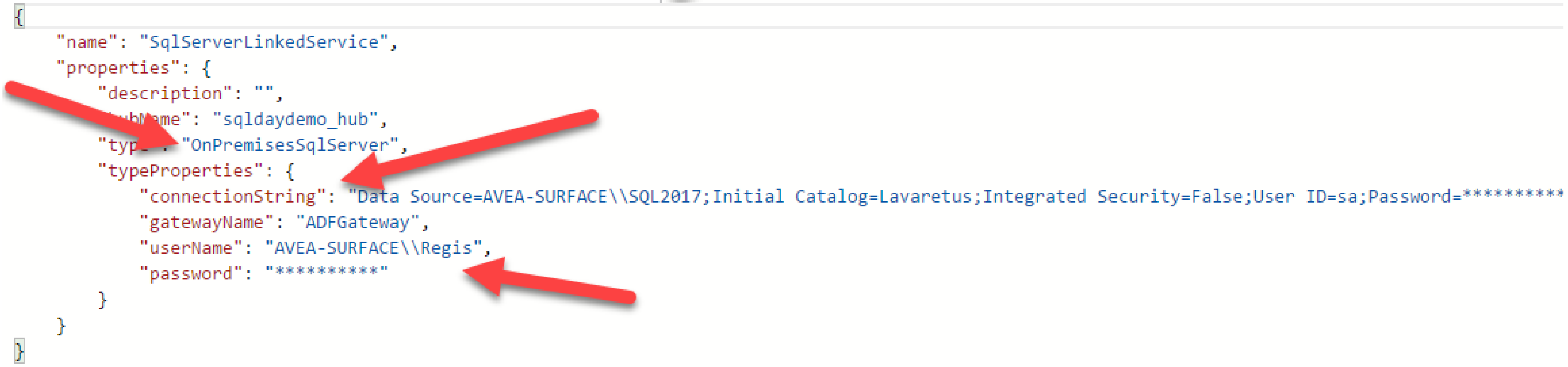
Address

Protocol

Authentication



```
{
    "name": "tblCapacitiesFromBlob",
    "properties": {
        "structure": [
            {
                "name": "LRNO",
                "type": "String"
            },
            {
                "name": "Bale",
                "type": "String"
            }
        ],
        "published": false,
        "type": "AzureBlob",
        "linkedServiceName": "StorageLinkedService",
        "typeProperties": {
            "folderPath": "adfpoc/",
            "format": {
```



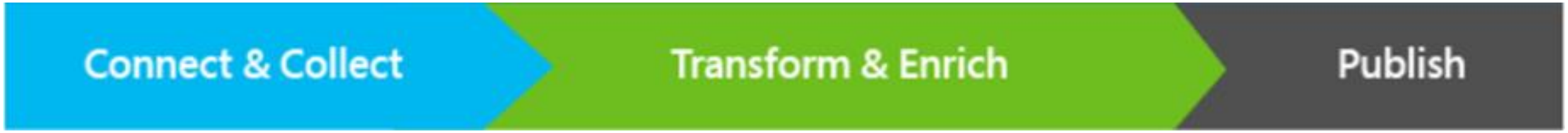
```
{
    "name": "SqlServerLinkedService",
    "properties": {
        "description": "",
        "hubName": "sqldaydemo_hub",
        "type": "OnPremisesSqlServer",
        "typeProperties": {
            "connectionString": "Data Source=AVEA-SURFACE\\SQL2017;Initial Catalog=Lavaretus;Integrated Security=False;User ID=sa;Password=**********
            "gatewayName": "ADFGateway",
            "userName": "AVEA-SURFACE\\Regis",
            "password": "**********"
        }
    }
}
```

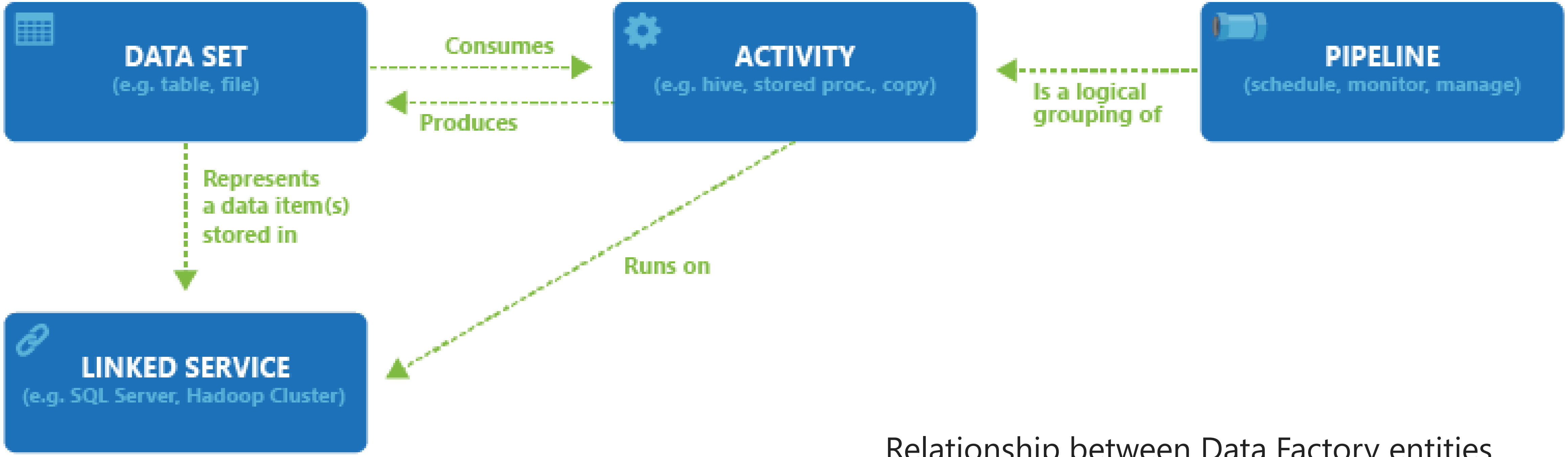
```
        "policy": {}
    }
}
```

Pipeline = logical grouping of activities

Data movement activities

Data transformation activities





DATA SET
(e.g. table, file)

Consumes

Produces

ACTIVITY
(e.g. hive, stored proc., copy)

Is a logical grouping of

PIPELINE
(schedule, monitor, manage)

Represents a data item(s) stored in

Runs on

LINKED SERVICE
(e.g. SQL Server, Hadoop Cluster)

Relationship between Data Factory entities

2 types of Activities:

Copy & Transform

Data Movement Activities

All are Sources / Sink are orange

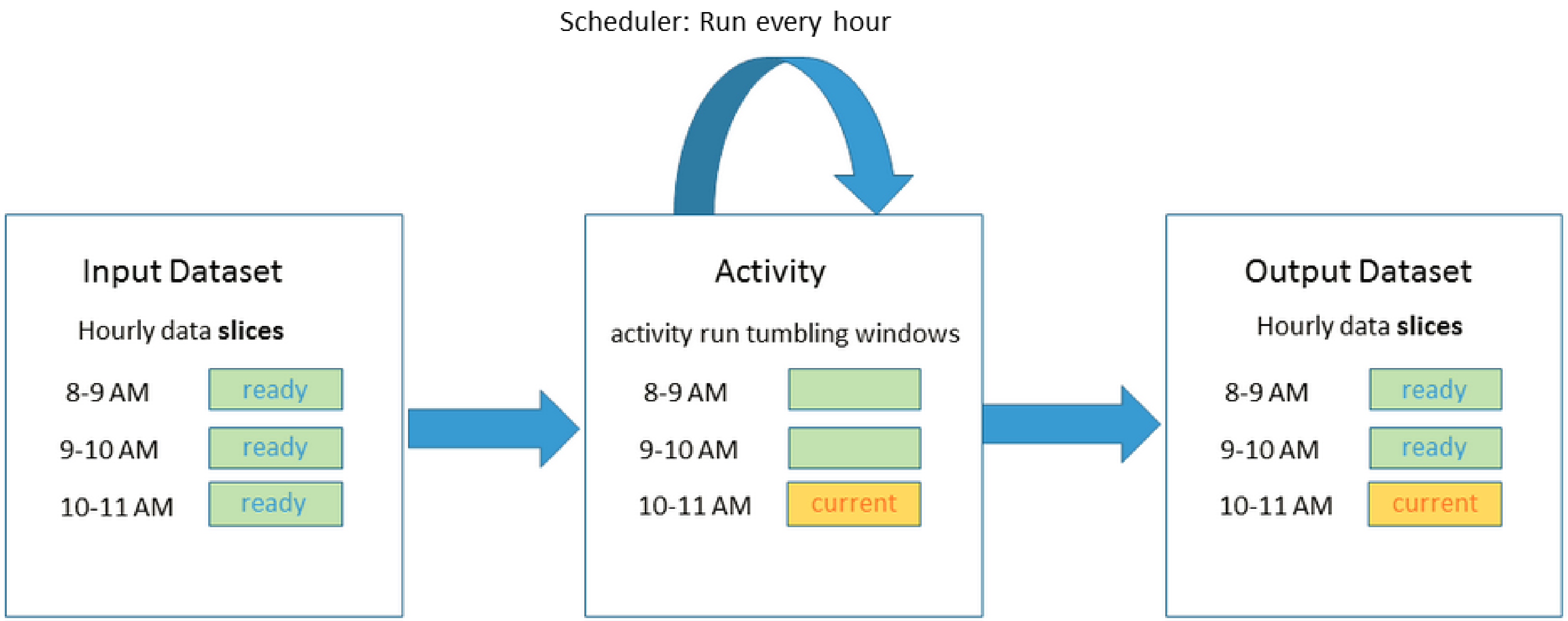
- Azure Blob storage
- Azure Data Lake Store
- Azure SQL Database
- Azure SQL Data Warehouse
- Azure Table storage
- Azure ~~Document~~ Cosmos DB
- Azure Search Index
- SQL Server*
- Oracle*
- MySQL*
- DB2*
- Teradata*
- PostgreSQL*
- Sybase*
- Cassandra*
- MongoDB*
- Amazon Redshift
- File System*
- HDFS*
- Amazon S3
- FTP
- Salesforce
- Generic ODBC*
- Generic OData
- Web Table (table from HTML)
- GE Historian*

Data Transform Activities

Transformation	HDInsight [Hadoop]
Hive	HDInsight [Hadoop]
Pig	HDInsight [Hadoop]
MapReduce	HDInsight [Hadoop]
Hadoop Streaming	Azure VM
Machine Learning activities: Batch Execution and Update	
Resource	Azure SQL, Azure SQL Data Warehouse, or SQL Server
Stored Procedure	Azure Data Lake Analytics
Data Lake Analytics U-SQL	HDInsight [Hadoop] or Azure Batch
DotNet	

Recurring schedule

Data slices



Scheduler: Run every hour

Input Dataset

Hourly data **slices**

8-9 AM	ready
9-10 AM	ready
10-11 AM	ready

Activity

activity run tumbling windows

8-9 AM	
9-10 AM	
10-11 AM	current

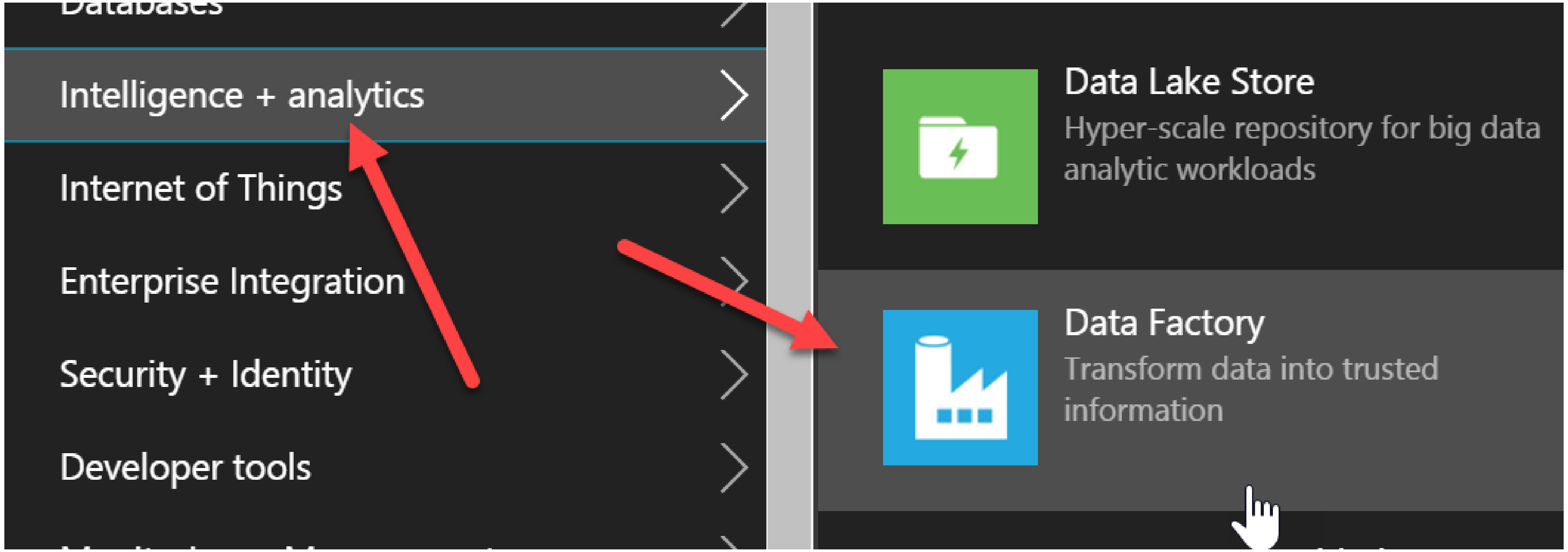
Output Dataset

Hourly data **slices**

8-9 AM	ready
9-10 AM	ready
10-11 AM	current

Transformation	**Compute Environment**
Hive	HDInsight [Hadoop]
Pig	HDInsight [Hadoop]
MapReduce	HDInsight [Hadoop]
Hadoop Streaming	HDInsight [Hadoop]
Machine Learning activities: Batch Execution and Update Resource	Azure VM
Stored Procedure	Azure SQL, Azure SQL Data Warehouse, or SQL Server
Data Lake Analytics U-SQL	Azure Data Lake Analytics
	HDInsight [Hadoop] or Azure Batch

PORTAL.AZURE.COM



Data Lake Store
Hyper-scale repository for big data analytic workloads

Data Factory
Transform data into trusted information

Linked Service represents

Data Stores = Source or Sink

Compute resource = Data transformation

Data Stores

Contain credentials and connection information for Sources and Destinations.

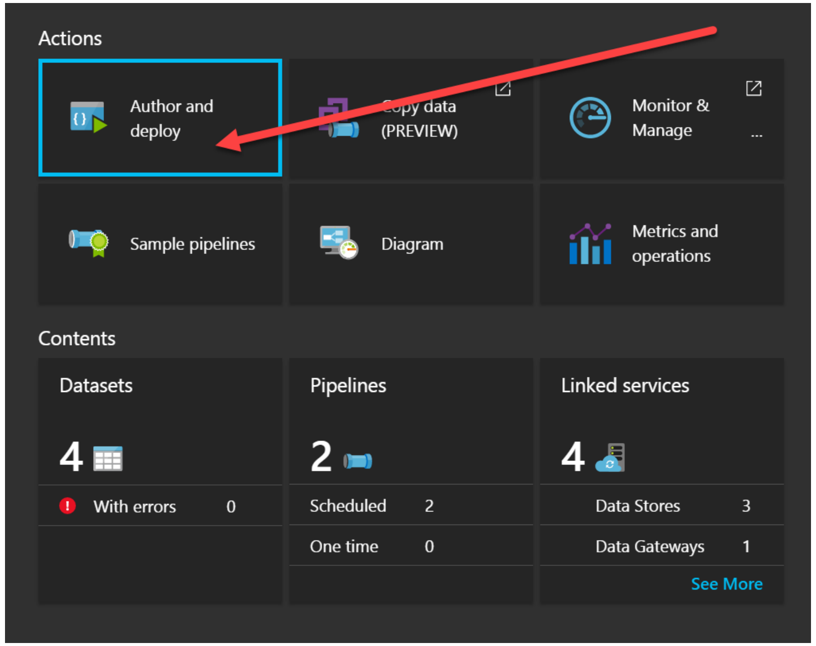
An On Premises Data Store **MUST** reference a Data Gateway

Dataset

Data structure in the data store

Building a pipeline → Azure Portal → Visual Studio → PowerShell → ARM Template → REST API

Demos



Actions

- Author and deploy
- Copy data (PREVIEW)
- Monitor & Manage
- Sample pipelines
- Diagram
- Metrics and operations

Contents

Datasets

4

With errors 0

Pipelines

2

Scheduled 2

One time 0

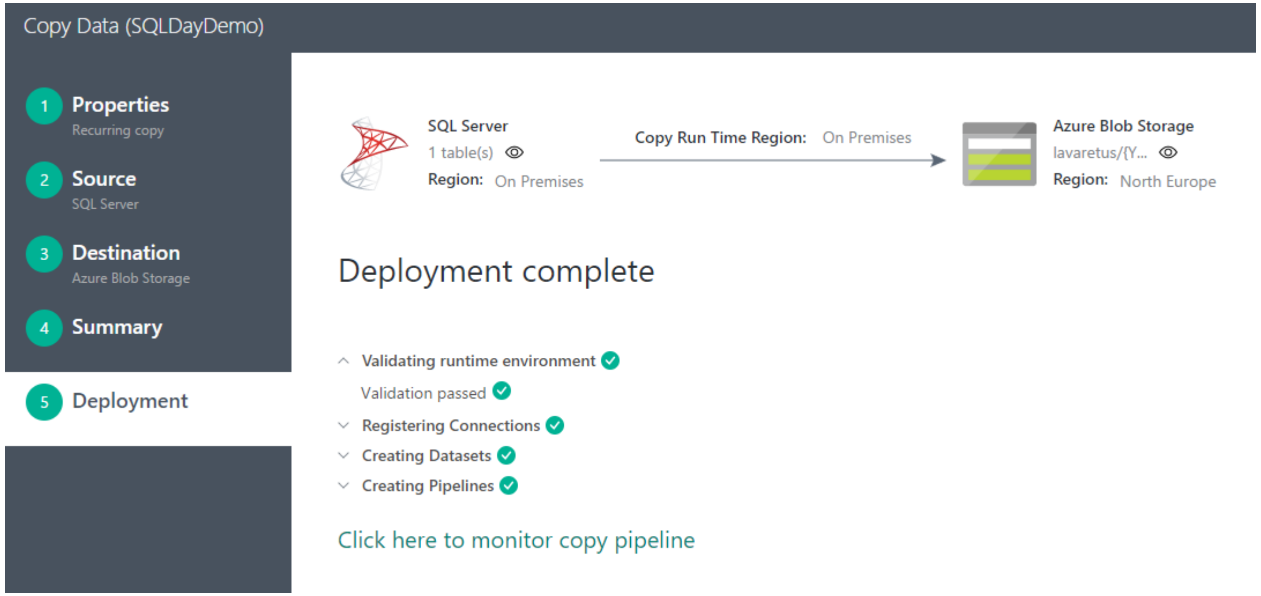
Linked services

4

Data Stores 3

Data Gateways 1

See More



Copy Data (SQLDayDemo)

1 **Properties** Recurring copy

2 **Source** SQL Server

3 **Destination** Azure Blob Storage

4 **Summary**

5 Deployment

SQL Server
1 table(s)
Region: On Premises

Copy Run Time Region: On Premises

Azure Blob Storage
lavaretus/{Y...
Region: North Europe

Deployment complete

- Validating runtime environment ✓
  - Validation passed ✓
- Registering Connections ✓
- Creating Datasets ✓
- Creating Pipelines ✓

Click here to monitor copy pipeline

Building a pipeline

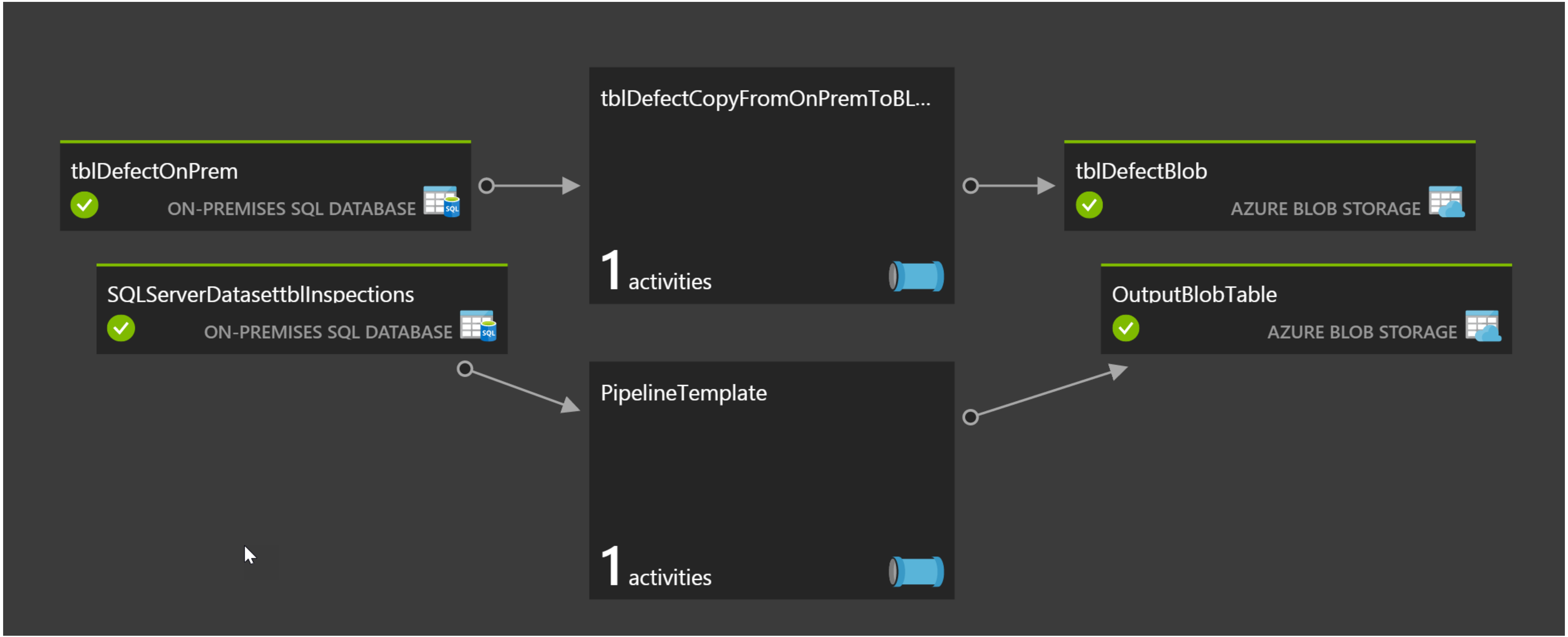
Azure Portal

Visual Studio

PowerShell

ARM Template

REST API



Building a pipeline

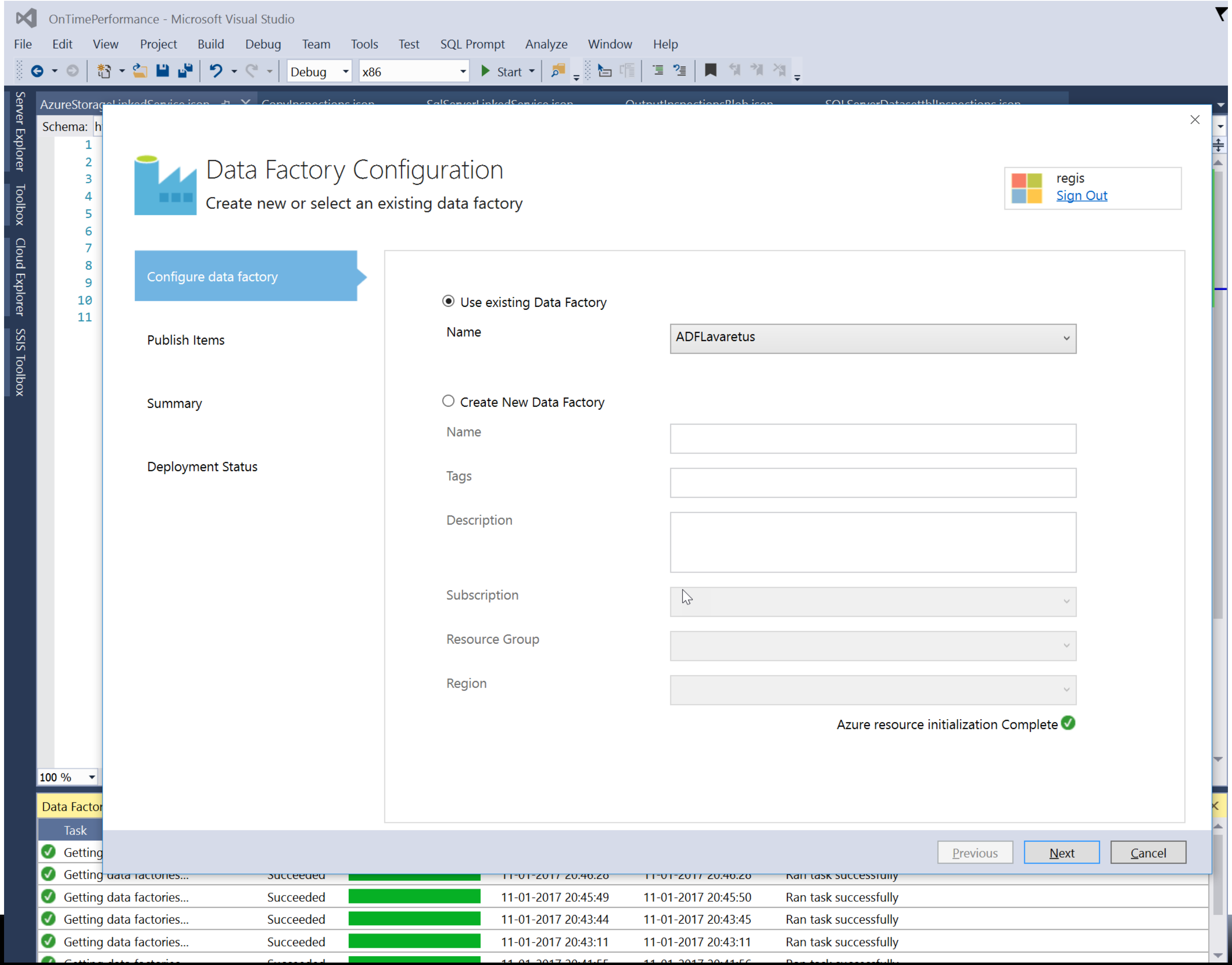
Azure Portal

Visual Studio

PowerShell

ARM Template

REST API



Data Factory Configuration

Create new or select an existing data factory

regis
Sign Out

Configure data factory

Publish Items

Summary

Deployment Status

● Use existing Data Factory

Name: ADFLavaretus

○ Create New Data Factory

Name

Tags

Description

Subscription

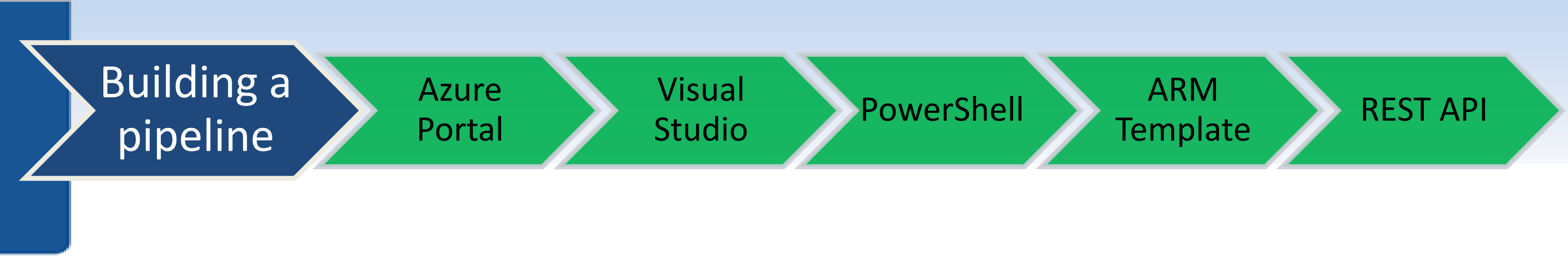
Resource Group

Region

Azure resource initialization Complete ✓

Previous | Next | Cancel

✓ Getting					
✓ Getting data factories...	Succeeded	🟩	11-01-2017 20:46:28	11-01-2017 20:46:28	Ran task successfully
✓ Getting data factories...	Succeeded	🟩	11-01-2017 20:45:49	11-01-2017 20:45:50	Ran task successfully
✓ Getting data factories...	Succeeded	🟩	11-01-2017 20:43:44	11-01-2017 20:43:45	Ran task successfully
✓ Getting data factories...	Succeeded	🟩	11-01-2017 20:43:11	11-01-2017 20:43:11	Ran task successfully


Building a pipeline

Azure Portal

Visual Studio

PowerShell

ARM Template

REST API

PowerShell

Get-AzureRmDataFactoryDataset

Get-AzureRmDataFactoryGateway

Get-AzureRmDataFactoryHub

Get-AzureRmDataFactoryLinkedService

Get-AzureRmDataFactoryPipeline

Get-AzureRmDataFactoryRun

Get-AzureRmDataFactorySlice

Azure Resource Manager Template

More JSON !!

REST API

CURL Tool with REST

Create Web Application in AAD

Assign to Data Factory Contributor Role

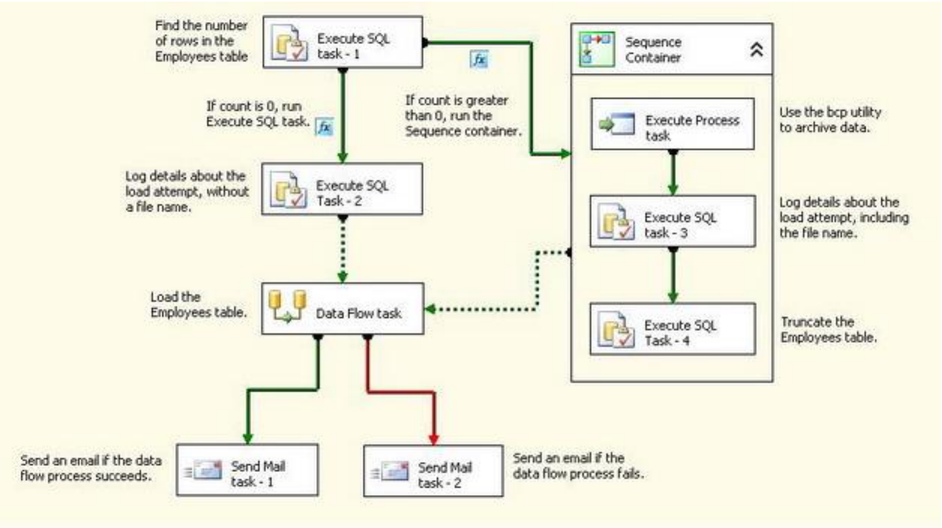
Use CURL to communicate with Web Application

```
$cmd = {.\curl.exe -X PUT -H "Authorization: Bearer $accessToken" -H "Content-Type: application/json" --data “@azurestoragelinkedservice.json” https://management.azure..

$results = Invoke-Command -scriptblock $cmd;
```

Data Extract, Transformation and Loading tool

Born in 2005 – enterprise ready but still room for new features

Part of SQL Server license and installation


Find the number of rows in the Employees table
Execute SQL task - 1
Sequence Container
If count is 0, run Execute SQL task.
If count is greater than 0, run the Sequence container.
Execute Process task
Use the bcp utility to archive data.
Log details about the load attempt, without a file name.
Execute SQL Task - 2
Execute SQL task - 3
Log details about the load attempt, including the file name.
Load the Employees table.
Data Flow task
Execute SQL Task - 4
Truncate the Employees table.
Send an email if the data flow process succeeds.
Send Mail task - 1
Send Mail task - 2
Send an email if the data flow process fails.

SSIS 101

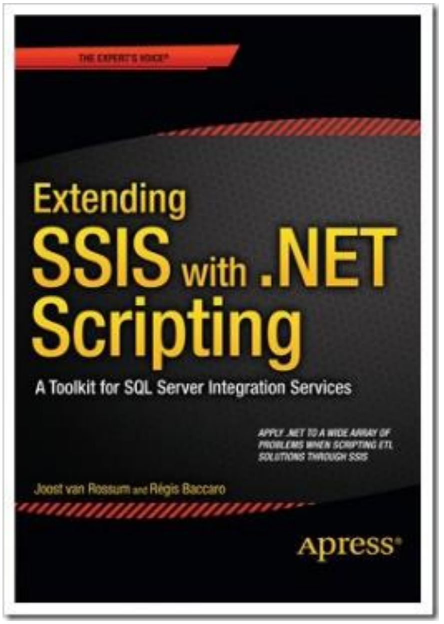
What is SSIS for?

Benefits

Rich development tool

Many built-in transformations

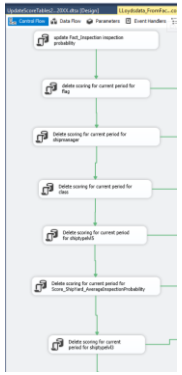
Extensible with scripts

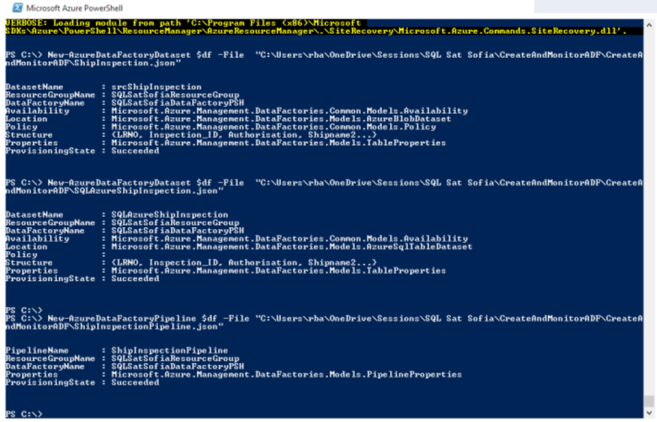

THE EXPERT'S VOICE®
Extending
SSIS with .NET
Scripting
A Toolkit for SQL Server Integration Services
APPLY .NET TO A WIDE ARRAY OF PROBLEMS WHEN SCRIPTING ETL SOLUTIONS THROUGH SSIS
Joost van Rossum and Régis Baccaro
Apress®

- Development tool
- Administration tool
- Data source & destinations
- Data transformations
- Price
- Error handling
- Deployment
- Monitoring
- Security
- Technology
- Requirements
- Big Data compatibility

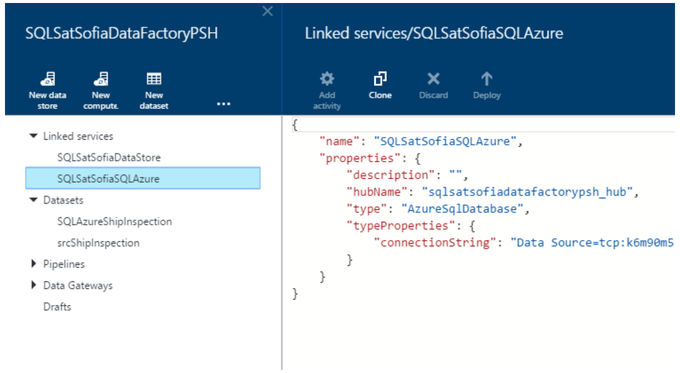
SSIS	ADF
SSDT	Azure portal ADF Editor
Free	PowerShell
	JSON Scripts
	Visual Studio

	SSIS	ADF
Standalone tool	Yes	Yes
Powerful GUI	Yes	No
Available	Free	Free
Prerequisite	SQL Server	Azure Subscription


UpdateScoreTables2...200X.dtsx [Design]
LLoydsdata_FromFac...
Control Flow
Data Flow
Parameters
Event Handlers
update Fact_Inspection inspection probability
delete scoring for current period for flag
Delete scoring for current period for shipmanager
Delete scoring for current period for class
Delete scoring for current period for shiptypelvl5
Delete scoring for current period for Score_ShipYard_AverageInspectionProbability
Delete scoring for current period for shiptypelvl3


Microsoft Azure PowerShell

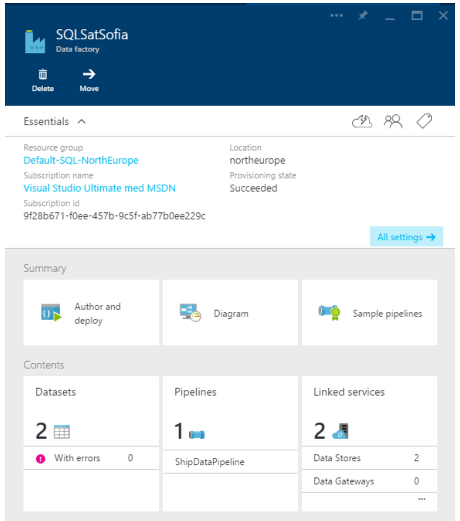
```
VERBOSE: Loading module from path 'C:\Program Files (x86)\Microsoft
SDKs\Azure\PowerShell\ResourceManager\AzureResourceManager\.\SiteRecovery\Microsoft.Azure.Commands.SiteRecovery.dll'.

PS C:\> New-AzureDataFactoryDataset $df -File  "C:\Users\rba\OneDrive\Sessions\SQL Sat Sofia\CreateAndMonitorADF\CreateA
ndMonitorADF\ShipInspection.json"


DatasetName       : srcShipInspection
ResourceGroupName : SQLSatSofiaResourceGroup
DataFactoryName   : SQLSatSofiaDataFactoryPSH
Availability      : Microsoft.Azure.Management.DataFactories.Common.Models.Availability
Location          : Microsoft.Azure.Management.DataFactories.Models.AzureBlobDataset
Policy            : Microsoft.Azure.Management.DataFactories.Common.Models.Policy
Structure         : {LRNO, Inspection_ID, Authorisation, Shipname2...}
Properties        : Microsoft.Azure.Management.DataFactories.Models.TableProperties
ProvisioningState : Succeeded



PS C:\> New-AzureDataFactoryDataset $df -File  "C:\Users\rba\OneDrive\Sessions\SQL Sat Sofia\CreateAndMonitorADF\CreateA
ndMonitorADF\SQLAzureShipInspection.json"


DatasetName       : SQLAzureShipInspection
ResourceGroupName : SQLSatSofiaResourceGroup
DataFactoryName   : SQLSatSofiaDataFactoryPSH
Availability      : Microsoft.Azure.Management.DataFactories.Common.Models.Availability
Location          : Microsoft.Azure.Management.DataFactories.Models.AzureSqlTableDataset
Policy            :
Structure         : {LRNO, Inspection_ID, Authorisation, Shipname2...}
Properties        : Microsoft.Azure.Management.DataFactories.Models.TableProperties
ProvisioningState : Succeeded



PS C:\>
PS C:\> New-AzureDataFactoryPipeline $df -File "C:\Users\rba\OneDrive\Sessions\SQL Sat Sofia\CreateAndMonitorADF\CreateA
ndMonitorADF\ShipInspectionPipeline.json"


PipelineName      : ShipInspectionPipeline
ResourceGroupName : SQLSatSofiaResourceGroup
DataFactoryName   : SQLSatSofiaDataFactoryPSH
Properties        : Microsoft.Azure.Management.DataFactories.Models.PipelineProperties
ProvisioningState : Succeeded



PS C:\>
```



```
{
    "name": "SQLSatSofiaSQLAzure",
    "properties": {
        "description": "",
        "hubName": "sqlsatsofiadatafactorypsh_hub",
        "type": "AzureSqlDatabase",
        "typeProperties": {
            "connectionString": "Data Source=tcp:k6m90m5
        }
    }
}
```

Comparing SSIS and ADF

SSIS	ADF
SSMS	Azure portal
PowerShell	PowerShell w/ADF Cmdlets



SQLSatSofia
Data factory

Essentials

Resource group
Default-SQL-NorthEurope

Location
northeurope

Subscription name
Visual Studio Ultimate med MSDN

Provisioning state
Succeeded

Subscription id
9f28b671-f0ee-457b-9c5f-ab77b0ee229c

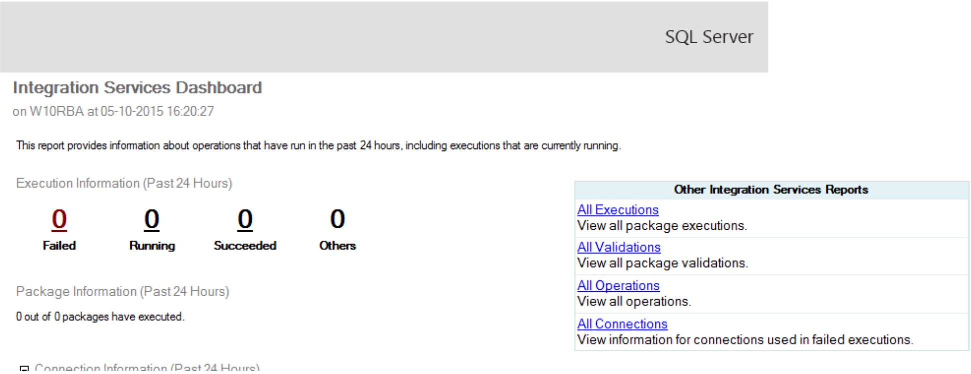
Summary

Contents

Datasets
2
With errors 0

Pipelines
1
ShipDataPipeline

Linked services
2
Data Stores 2
Data Gateways 0



SQL Server

Integration Services Dashboard

on W10RBA at 05-10-2015 16:20:27

This report provides information about operations that have run in the past 24 hours, including executions that are currently running.

Execution Information (Past 24 Hours)

0	0	0	0
Failed	Running	Succeeded	Others

Package Information (Past 24 Hours)

0 out of 0 packages have executed.

Connection Information (Past 24 Hours)

Other Integration Services Reports
All Executions View all package executions.
All Validations View all package validations.
All Operations View all operations.
All Connections View information for connections used in failed executions.

Orchestration

- In the cloud
  - High frequency: €0,51 per activity/month
  - Low frequency: €0,84 per activit
- On-prem with Data Mgt Ga
  - High frequenc nth
  - Low fre ivity/month

First five are free

20% discount for more than 100 activities

Data movement

- Between Cloud data stores: €0,21/hour
- When on-prem is involved: €0,08/hour

Inactive Pipelines:€0,67/month

Re-running activities

- In the Cloud: €1,13 per 1000 re-runs
- On-prem: €2,83 per 1000 re-runs

Inactive Pipelines:€0,67/month

Calculator

https://azure.microsoft.com/en-us/pricing/calculator/?service=data-factory

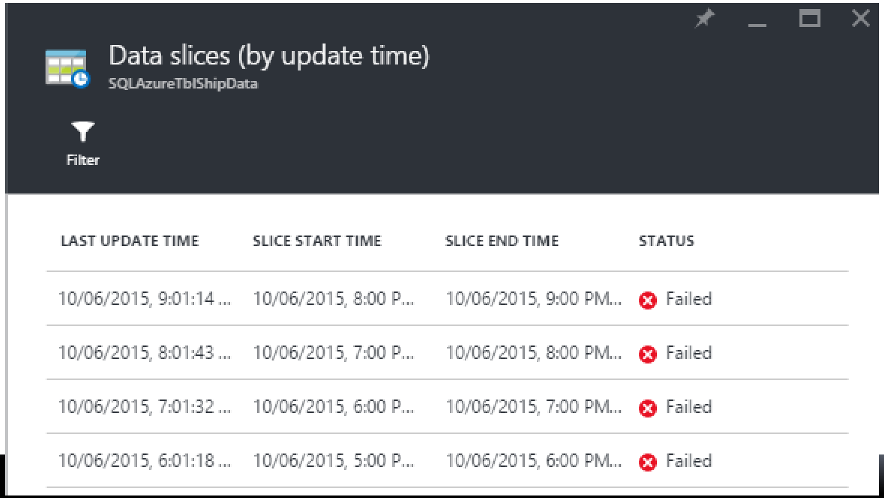
- Basic SSIS
    - Free = Express edition – Import & Export wizard
- Standard SSIS
    - Standard (and BI editions)
- Enterprise SSIS (CDC & Advanced adapters)
    - Enterprise edition
- Still true with SQL Server 2016 SP1

	ADF	SSIS
Licensing		Yes
Pay for features		Yes
Pay per usage	Yes	

	ADF	SSIS
Azure environment	Yes	
Hardware setup		Yes
Software setup		Yes
Administration costs / Data center		Yes

- SSIS
    - MSDB
    - Project deployment
    - Package deployment
    - SSIS Catalog
- ADF
    - Power shell scripts
    - Automization w/ PS Scripts

	ADF	SSIS
Alerts		Yes
Error Loging	Yes	Yes
Error handling		Yes



Data slices (by update time)
SQLAzureTblShipData

LAST UPDATE TIME	SLICE START TIME	SLICE END TIME	STATUS
10/06/2015, 9:01:14 ...	10/06/2015, 8:00 P...	10/06/2015, 9:00 PM...	Failed
10/06/2015, 8:01:43 ...	10/06/2015, 7:00 P...	10/06/2015, 8:00 PM...	Failed
10/06/2015, 7:01:32 ...	10/06/2015, 6:00 P...	10/06/2015, 7:00 PM...	Failed
10/06/2015, 6:01:18 ...	10/06/2015, 5:00 P...	10/06/2015, 6:00 PM...	Failed

SSIS Logging

SSIS Catalog reports

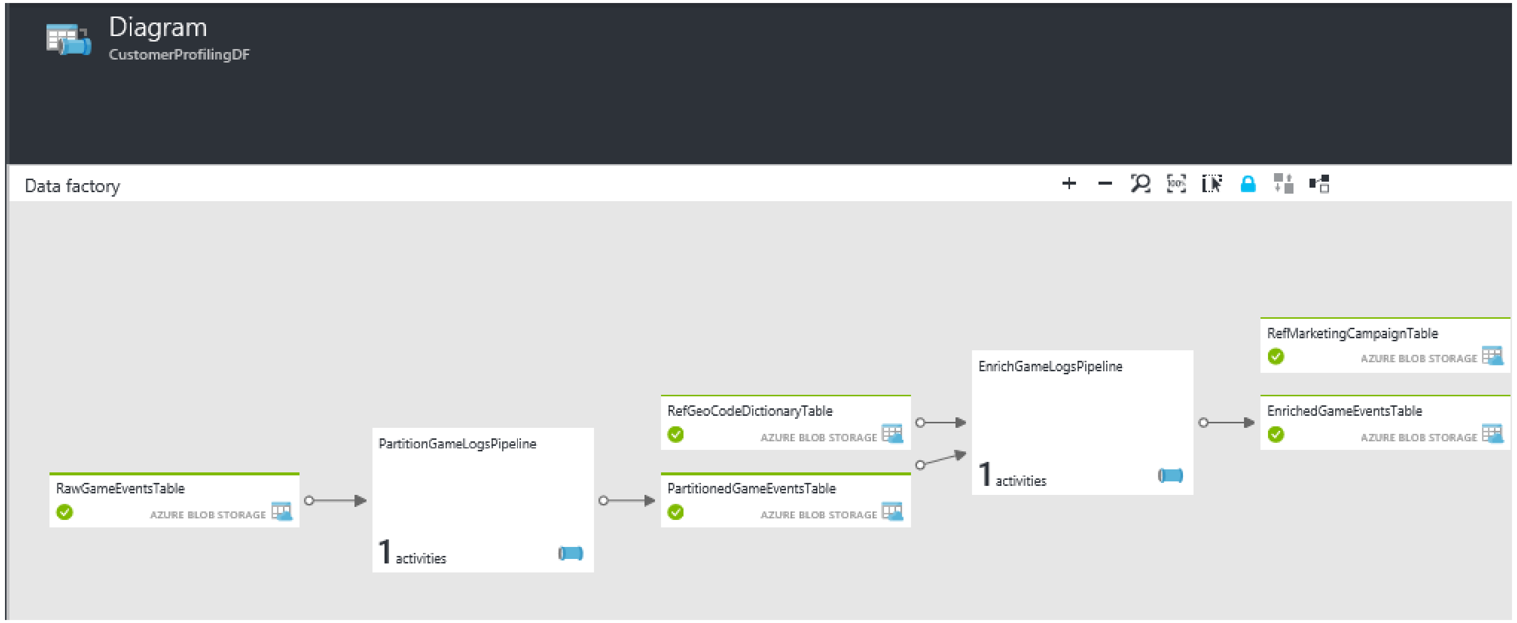
Diagram view

Drill through features

Capable GUI

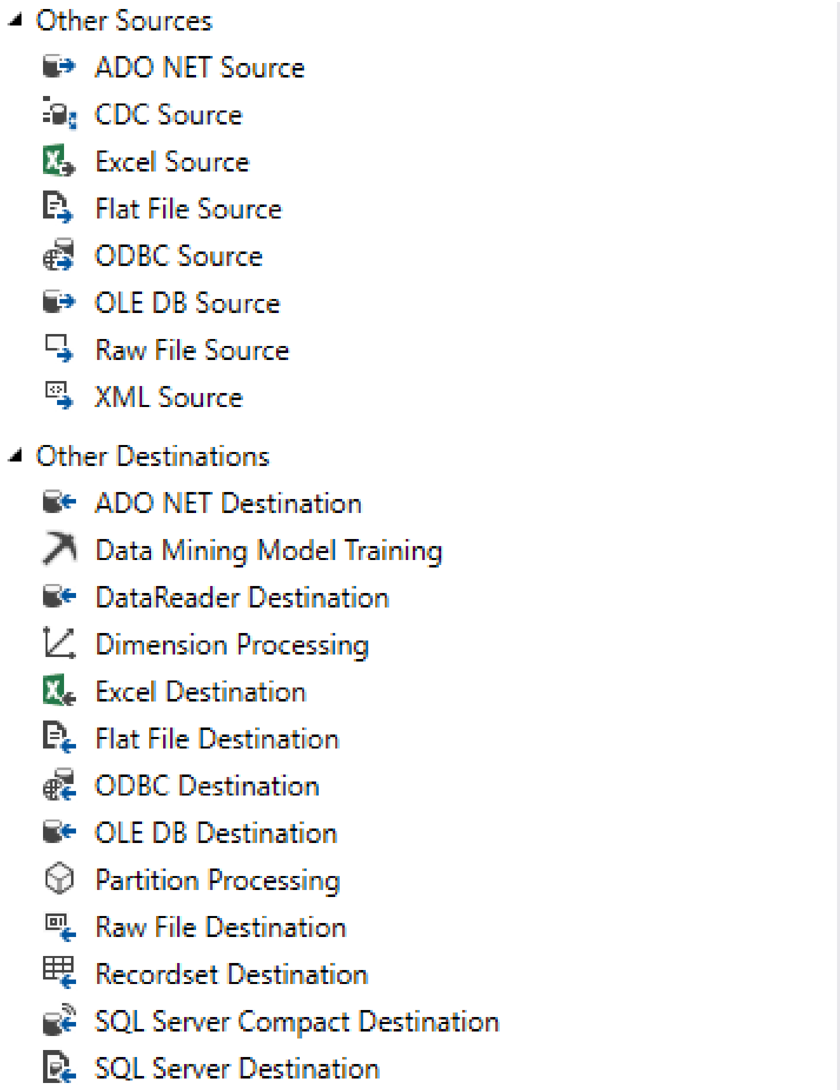
Data slice execution

Data lineage



	ADF	SSIS
Monitoring GUI	Yes	Yes
Drillthrough	Yes	Yes
Data slice	Yes	
Data lineage	Yes	

SQL Server
Oracle
SAP
Azure
Access
Sybase
PostGresSQL
FoxPro
SharePoint
WebService
.....

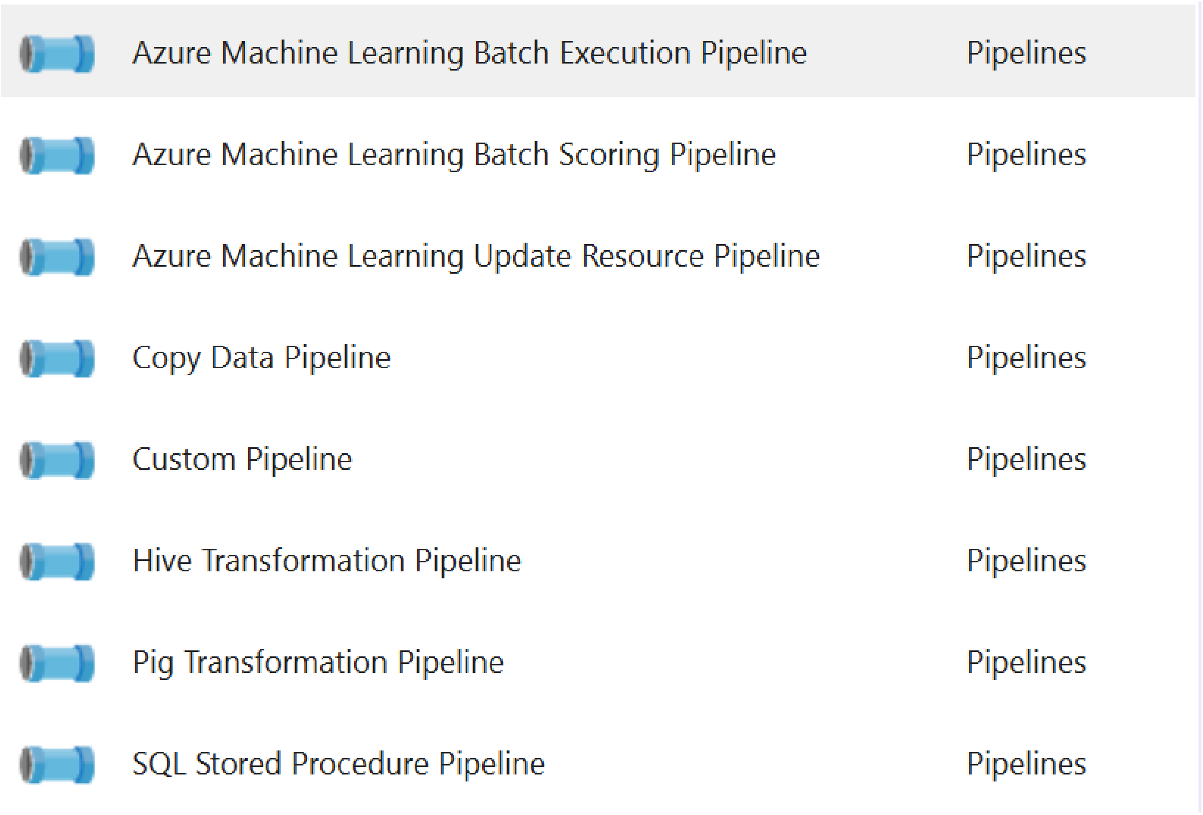


- Other Sources
  - ADO NET Source
  - CDC Source
  - Excel Source
  - Flat File Source
  - ODBC Source
  - OLE DB Source
  - Raw File Source
  - XML Source
- Other Destinations
  - ADO NET Destination
  - Data Mining Model Training
  - DataReader Destination
  - Dimension Processing
  - Excel Destination
  - Flat File Destination
  - ODBC Destination
  - OLE DB Destination
  - Partition Processing
  - Raw File Destination
  - Recordset Destination
  - SQL Server Compact Destination
  - SQL Server Destination

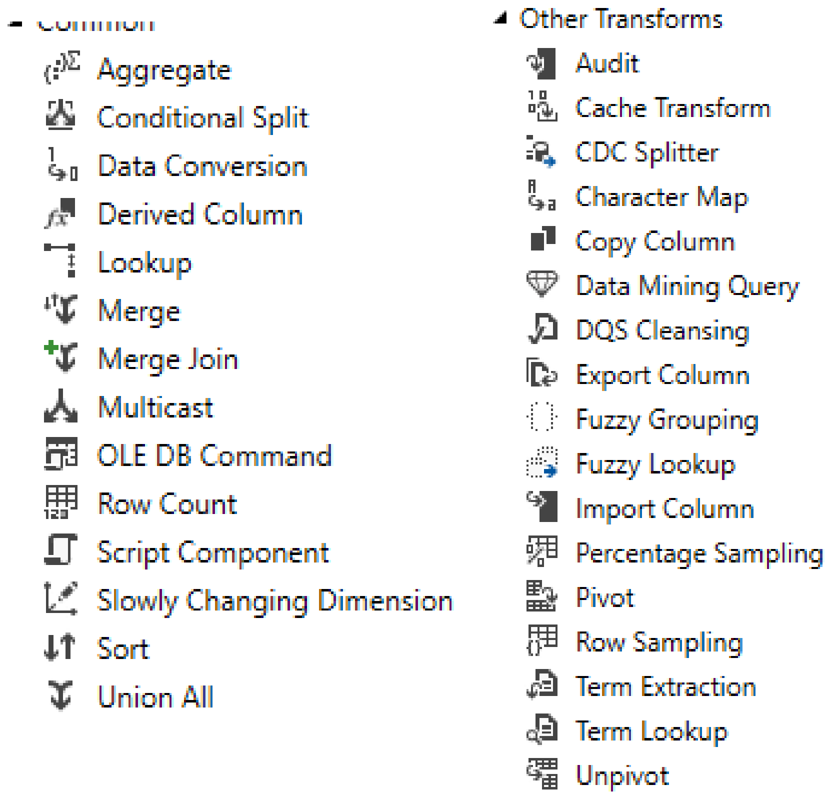
All are Sources / Sink are orange

Azure Blob storage
Azure Data Lake Store
Azure SQL Database
Azure SQL Data Warehouse
Azure Table storage
Azure DocumentDB
[Azure Search Index]
SQL Server*
Oracle*
MySQL*
DB2*
Teradata*
PostgreSQL*

Sybase*
Cassandra*
MongoDB*
Amazon Redshift
File System*
HDFS*
Amazon S3
FTP
Salesforce
Generic ODBC*
Generic OData
Web Table (table from HTML)
GE Historian*



Azure Machine Learning Batch Execution Pipeline	Pipelines
Azure Machine Learning Batch Scoring Pipeline	Pipelines
Azure Machine Learning Update Resource Pipeline	Pipelines
Copy Data Pipeline	Pipelines
Custom Pipeline	Pipelines
Hive Transformation Pipeline	Pipelines
Pig Transformation Pipeline	Pipelines
SQL Stored Procedure Pipeline	Pipelines



◢ Common

- Aggregate
- Conditional Split
- Data Conversion
- Derived Column
- Lookup
- Merge
- Merge Join
- Multicast
- OLE DB Command
- Row Count
- Script Component
- Slowly Changing Dimension
- Sort
- Union All

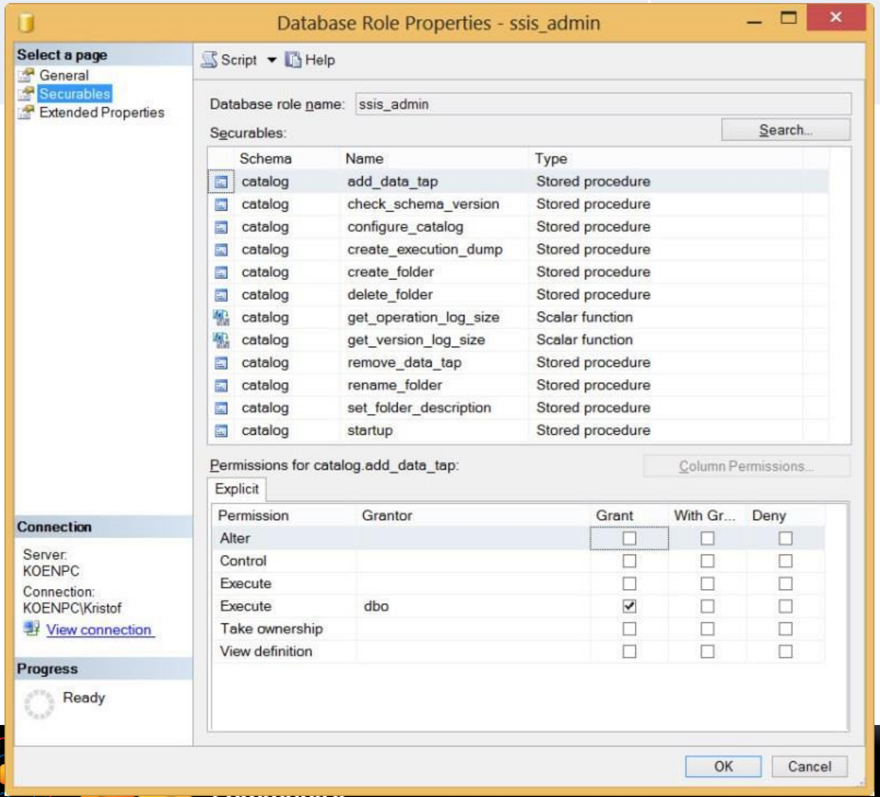
◢ Other Transforms

- Audit
- Cache Transform
- CDC Splitter
- Character Map
- Copy Column
- Data Mining Query
- DQS Cleansing
- Export Column
- Fuzzy Grouping
- Fuzzy Lookup
- Import Column
- Percentage Sampling
- Pivot
- Row Sampling
- Term Extraction
- Term Lookup
- Unpivot

	ADF	SSIS
Copy	Yes	Yes
C# custom transformations	Yes	Yes
Pig and Hive	Yes	Yes*
Azure ML Scoring	Yes	With scripting
Stored procedure	Yes	Yes
Built-in transformations		Yes

Comparing SSIS and ADF → Dev → Admin → Deployment → Monitoring → Source & Destinations → Security → Performance

	ADF	SSIS
Role based	Yes	Yes



Database Role Properties - ssis_admin

Database role name: ssis_admin

Securables:

Schema	Name	Type
catalog	add_data_tap	Stored procedure
catalog	check_schema_version	Stored procedure
catalog	configure_catalog	Stored procedure
catalog	create_execution_dump	Stored procedure
catalog	create_folder	Stored procedure
catalog	delete_folder	Stored procedure
catalog	get_operation_log_size	Scalar function
catalog	get_version_log_size	Scalar function
catalog	remove_data_tap	Stored procedure
catalog	rename_folder	Stored procedure
catalog	set_folder_description	Stored procedure
catalog	startup	Stored procedure

Permissions for catalog.add_data_tap:

Permission	Grantor	Grant	With Gr...	Deny
Alter		□	□	□
Control		□	□	□
Execute		□	□	□
Execute	dbo	☑	□	□
Take ownership		□	□	□
View definition		□	□	□

Connection

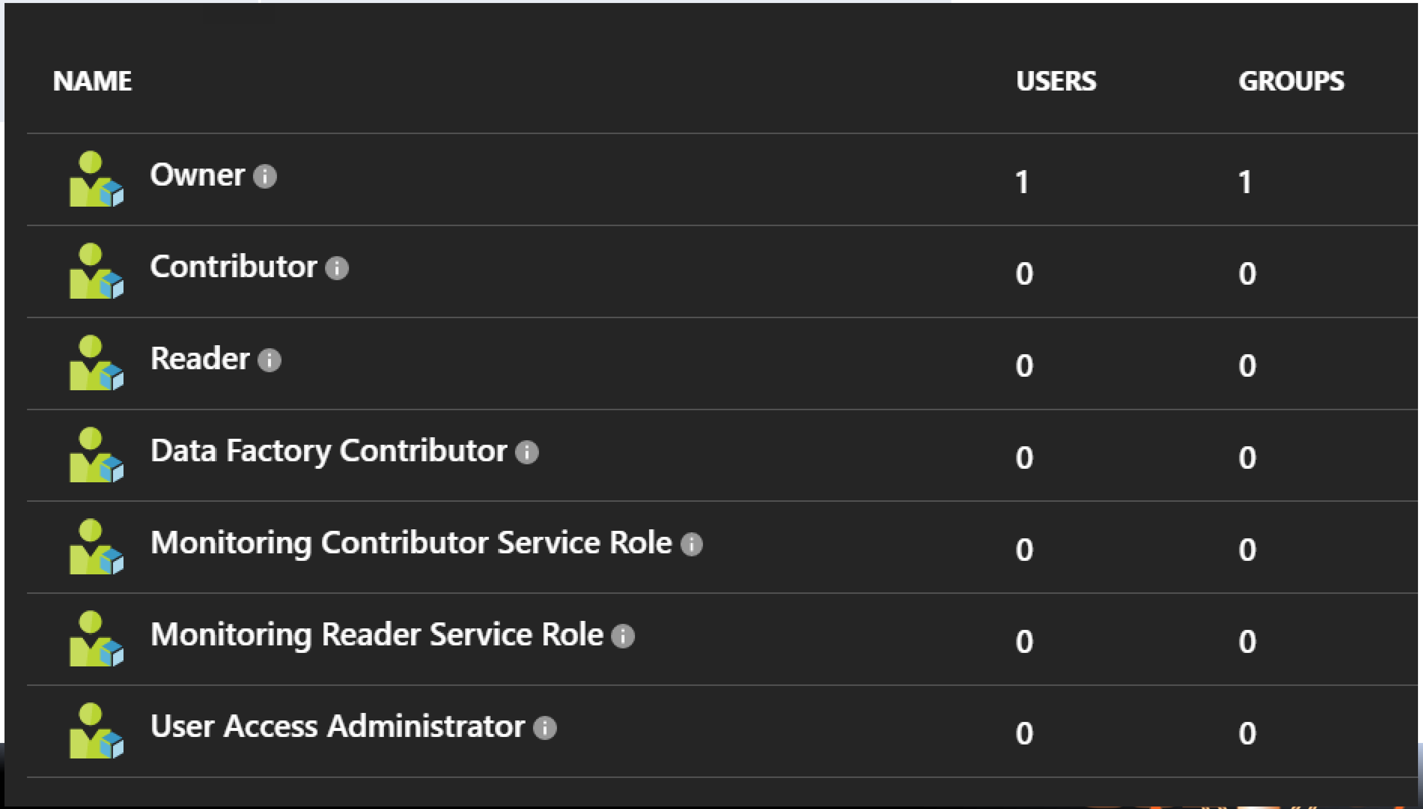
Server:
KOENPC

Connection:
KOENPC\Kristof

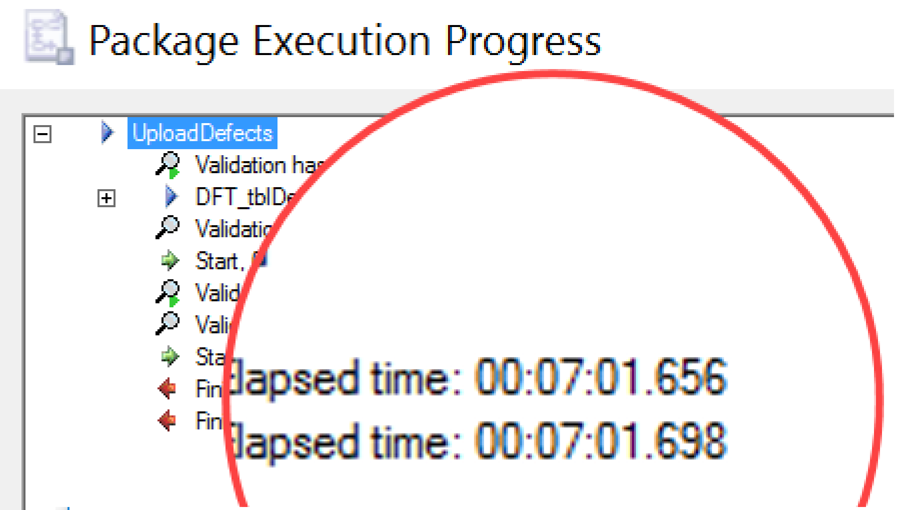
View connection

Progress

Ready

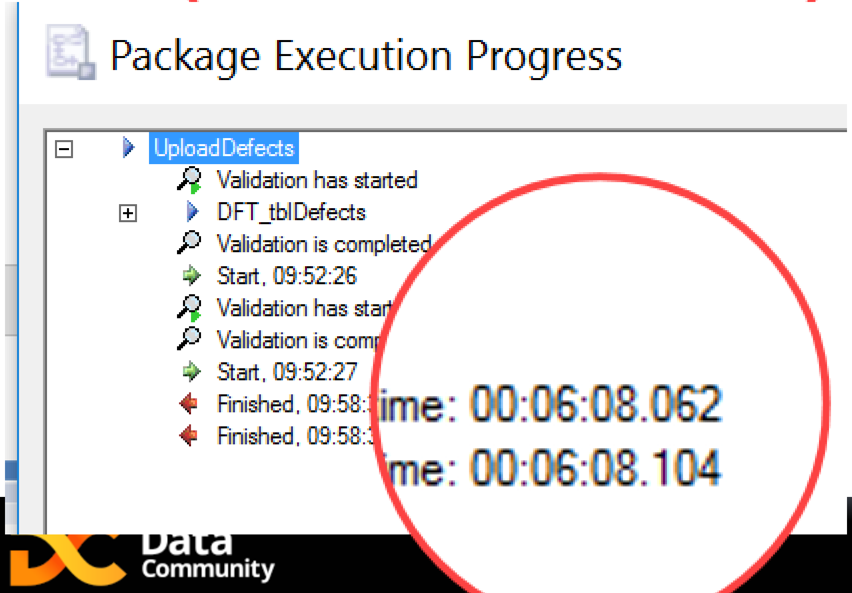


NAME	USERS	GROUPS
Owner	1	1
Contributor	0	0
Reader	0	0
Data Factory Contributor	0	0
Monitoring Contributor Service Role	0	0
Monitoring Reader Service Role	0	0
User Access Administrator	0	0



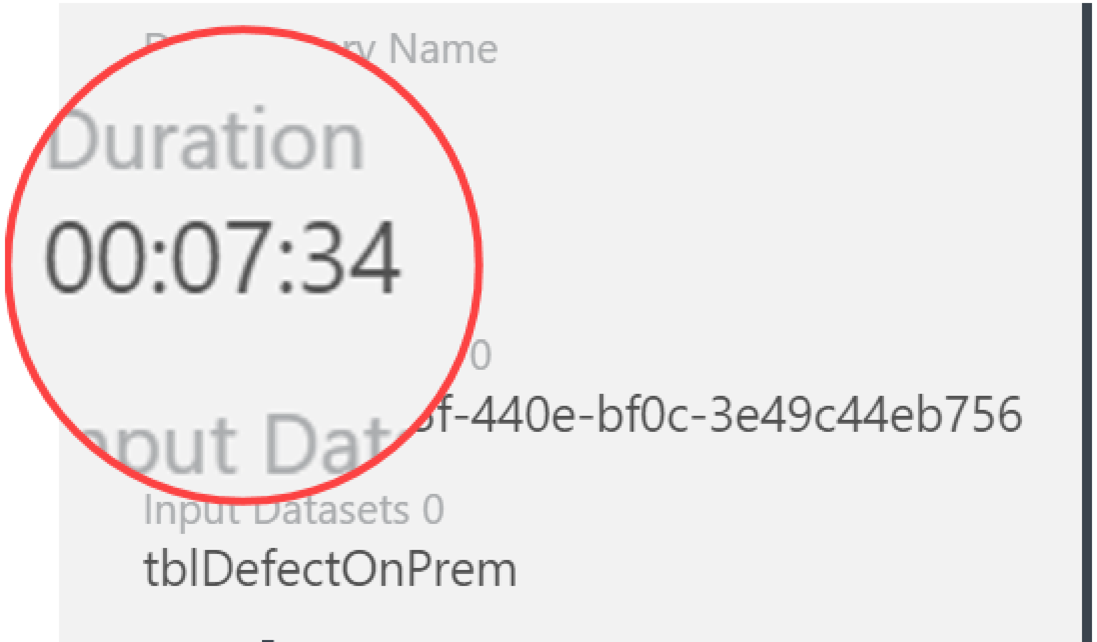
Package Execution Progress

```
⊟ ▶ UploadDefects
      Validation has
  ⊞   ▶ DFT_tblDe
      Validatio
      Start,
      Valid
      Vali
      Sta
      Fin  lapsed time: 00:07:01.656
      Fin  lapsed time: 00:07:01.698
```



Package Execution Progress

```
⊟ ▶ UploadDefects
      Validation has started
  ⊞   ▶ DFT_tblDefects
      Validation is completed
      Start, 09:52:26
      Validation has start
      Validation is comp
      Start, 09:52:27
      Finished, 09:58: ime: 00:06:08.062
      Finished, 09:58: me: 00:06:08.104
```



ry Name

Duration

00:07:34

0

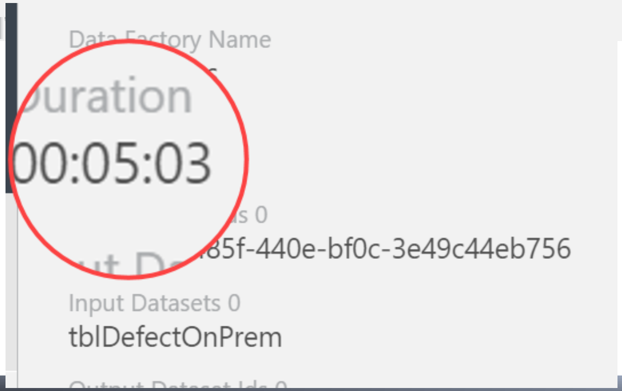
5f-440e-bf0c-3e49c44eb756

nput Dat

Input Datasets 0

tblDefectOnPrem

Outpu



Data Factory Name

uration

00:05:03

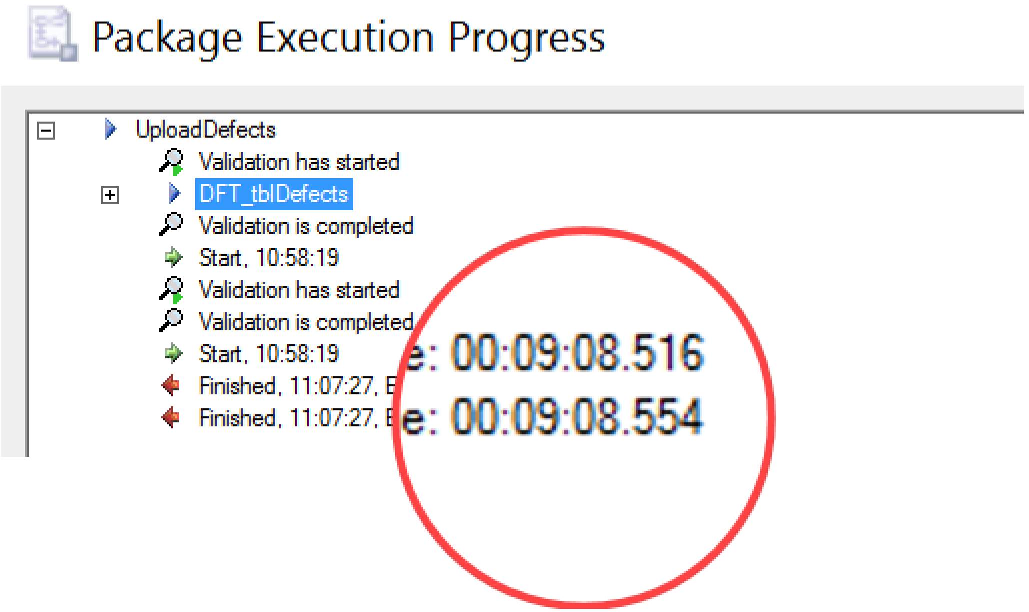
s 0

85f-440e-bf0c-3e49c44eb756

Input Datasets 0

tblDefectOnPrem

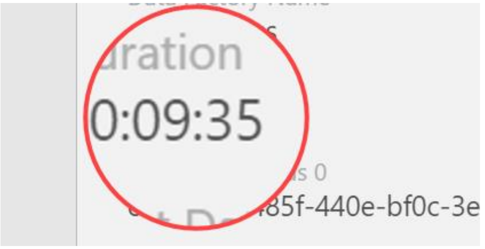
Output Dataset Ids 0



Package Execution Progress

```
UploadDefects
    Validation has started
    DFT_tblDefects
    Validation is completed
    Start, 10:58:19
    Validation has started
    Validation is completed
    Start, 10:58:19
    Finished, 11:07:27, E
    Finished, 11:07:27, E
```

```
e: 00:09:08.516
e: 00:09:08.554
```



```
uration
0:09:35
s 0
85f-440e-bf0c-3e
```

Conclusion

Not build for the same purpose

ADF still a V1 product

ADF great for Cloud Data integration for MS and Azure

Get benefits of both

- Hybrid SSIS and ADF
- Cloud based data movement, computing and monitoring
- On premises Data Transformations

Questions?

Contact me : regis@Baccaro.com

Twitter : @regbac

GOLD SPONSORS

Atos

Profisee®

Zerto

SILVER SPONSORS

PROMISE
GRUPA APN PROMISE

ELITMIND

ITMAGINATION
Imagine it. Done.

THEOBALD
SOFTWARE

BRONZE SPONSORS

it KONTRAKT®

JCommerce

IT
Krajowy Rejestr Długów

KSIĄŻKI IT Z CAŁEGO ŚWIATA
NOVATECH
www.novatech.com.pl

PEPCO

dbWatch
DATABASE CONTROL

querona

Data Scientist as a Service

UBS

VOLVO
Volvo Group IT

STRATEGIC PARTNER

Microsoft