Azure Data Factory – Part 1

with Paul Andrew

Paul Andrew

About Me

Microsoft Data Platform [MVP]

Business Intelligence Consultant at [Purple Frog Systems]

STEM Ambassador with [STEM Learning UK]

PASS Chapter Leader for [Microsoft Data Platform Group]

Speaker & part of the [SQL Relay] team

Speaker & helper at [SQL Bits]

Speaker & helper at various [SQL Saturday] 's

[Stack Overflow] Unsung Hero azure-data-factory

Many years' experience working with the complete on premises SQL Server stack in a variety of roles and industries. Now developing/consulting on hybrid business intelligence solutions using Microsoft Azure. Specialising in Data Factory, Data Lake Analytics, real-time data with the Internet of Things (IoT), Stream Analytics and Event Hubs.

linkedin.com/in/mrpaulandrew @mrpaulandrew paul@purplefrogsystems.com purplefrogsystems.com/paul

Session Agenda

Part 1

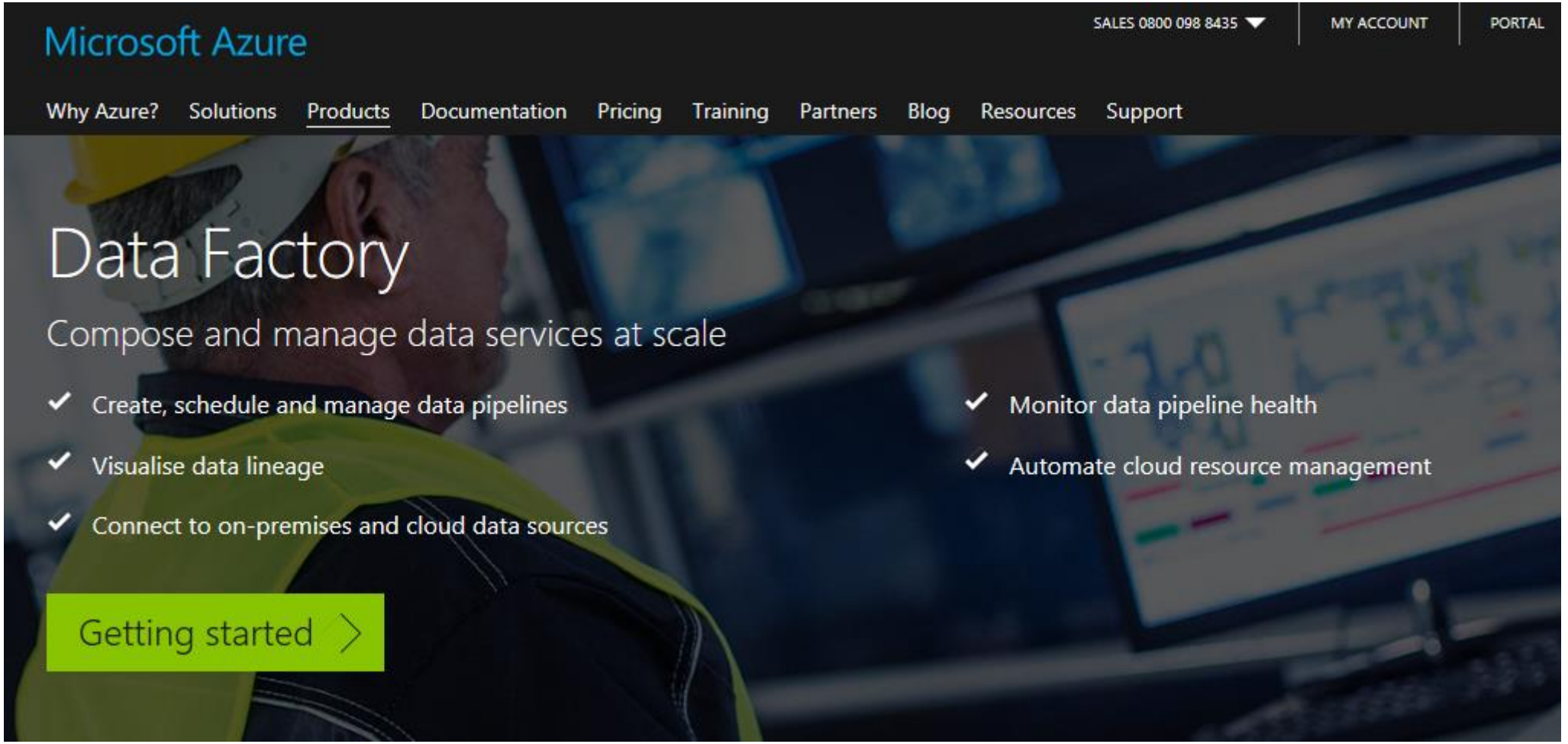
1. What is Azure Data Factory (ADF)?
   - What is an ADF Pipeline?
   - What is an ADF Dataset?
2. What is an ADF Activity?
3. What is an ADF Time Slice?
4. Developer tools for ADF.
5. The ADF Data Management Gateway.
6. Building a basic copy data factory.

Part 2

7. What is a ADF Custom Activity?
8. Creating a custom activity.
9. Using Visual Studio ADF config files.
10. Monitoring your ADF pipelines.
11. Q&A

What is Azure Data Factory (ADF)?

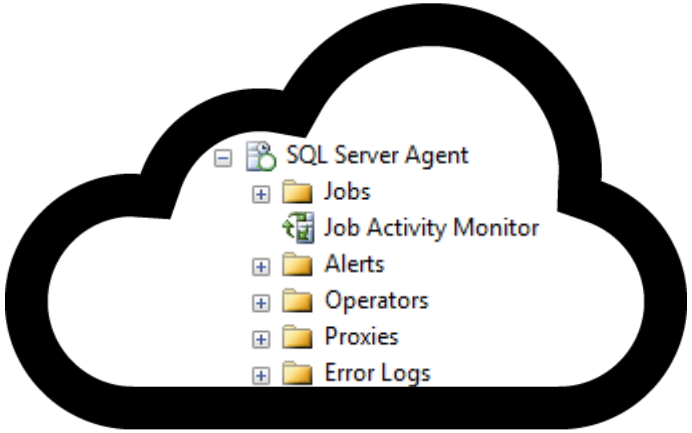
https://azure.microsoft.com/en-gb/services/data-factory/



What is Azure Data Factory (ADF)?

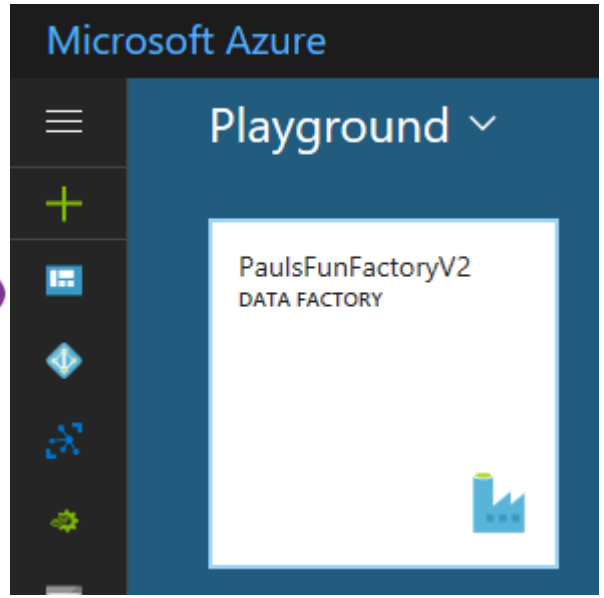
What it is…

What it is NOT…



What is an ADF Pipeline and Dataset?

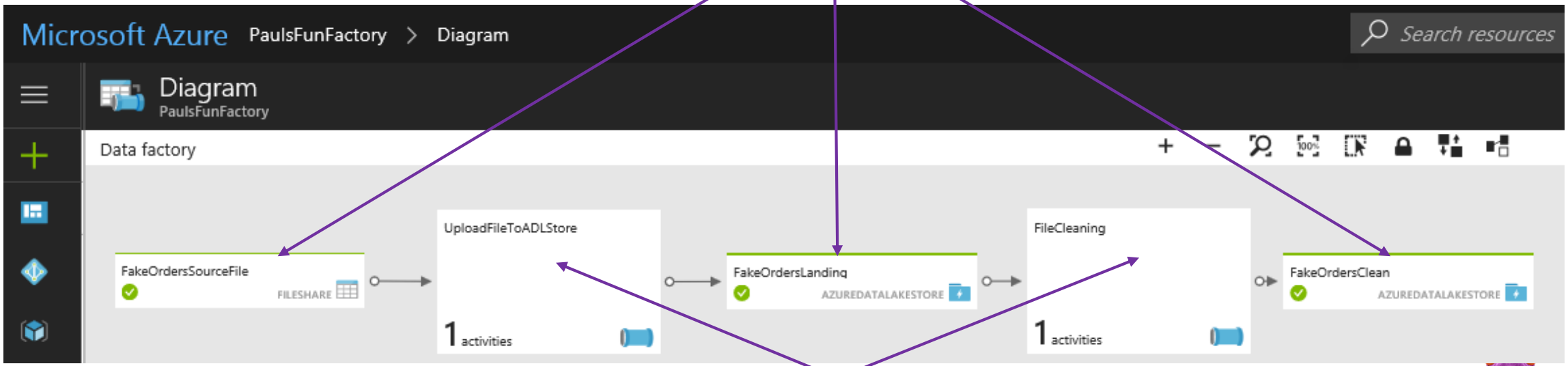
Blog post: https://www.purplefrogsystems.com/paul/2017/08/chaining-azure-data-factory-activities-and-datasets/



1x File

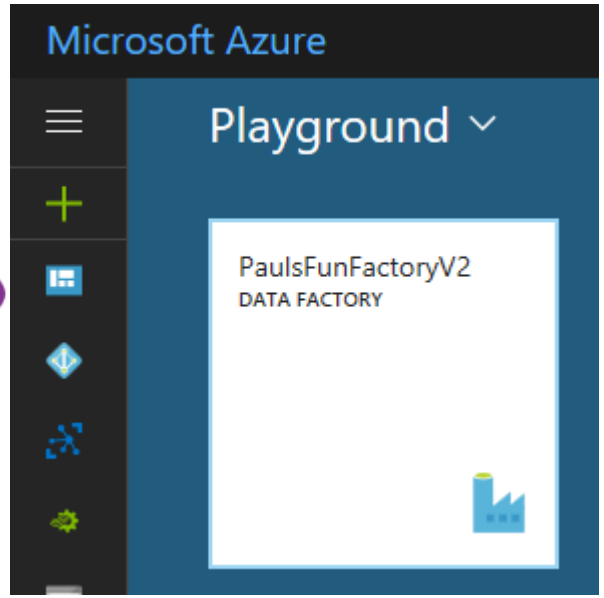
CSV

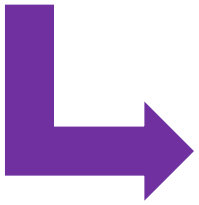
Datasets



Pipelines

What is an ADF Activity?

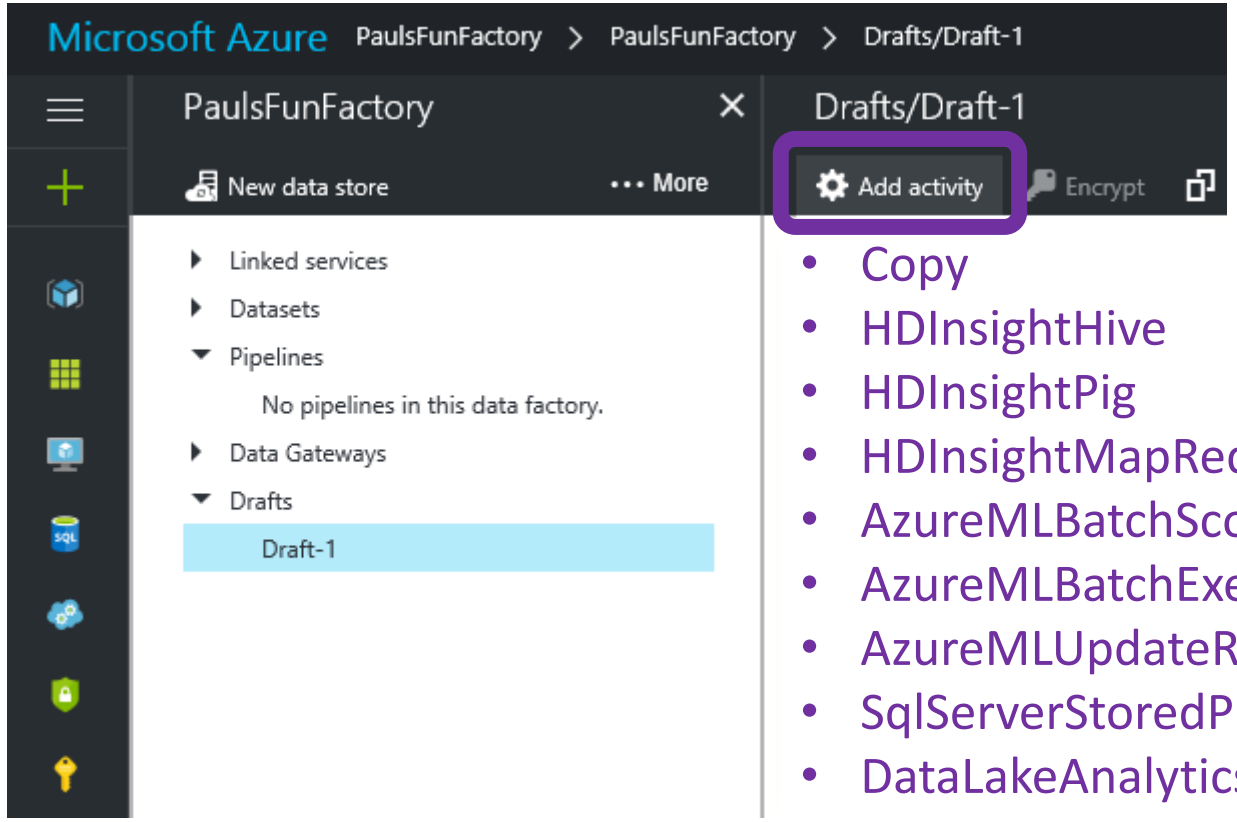




Charging:

https://azure.microsoft.com/en-gb/pricing/details/data-factory/

	Low Frequency (per month)	High Frequency (per month)
Cloud to Cloud	£0.45 per activity	£0.75 per activity
On Premises to Cloud	£1.12 per activity	£1.86 per activity



- Copy
- HDInsightHive
- HDInsightPig
- HDInsightMapReduce
- AzureMLBatchScoring
- AzureMLBatchExecution
- AzureMLUpdateResource
- SqlServerStoredProcedure
- DataLakeAnalyticsU-SQL

Meet JSON

DataLakeUploads.json* | ExcelExtractor.json

Schema: <No Schema Selected>

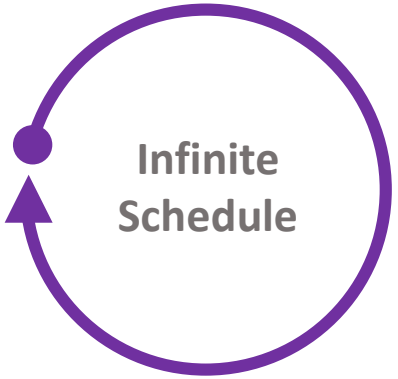
```
{
  "$schema": "http://datafactories.schema.management.azure.com/schemas/2015-09-01/Microsoft.DataFactory.Pipeline.json",
  "name": "DataLakeUploads",
  "properties": {
    "activities": [
      ..., //Do Stuff One
      ..., //Do Stuff Two
      ..., //Do Stuff Three
      ..., //Do Stuff Four
      ..., //Do Stuff Five
      ..., //Do Stuff Six
      ..., //Do Stuff Seven
      {
        "type": "Copy",
        "typeProperties": {
          "source": {
            "type": "FileSystemSource",
            "recursive": false
          },
          "sink": {
            "type": "FileSystemSink",
            "copyBehavior": "PreserveHierarchy",
            "writeBatchSize": 0,
            "writeBatchTimeout": "00:00:00"
          }
        },
        "inputs": [
          {
            "name": "File1OnPrem"
          }
        ],
        "outputs": [
          {
            "name": "File1ADFStore"
```

What is an ADF Time Slice?

What it is…

https://docs.microsoft.com/en-us/azure/data-factory/data-factory-scheduling-and-execution

What it is NOT…



Triggered Event *

* Unless you use another service, like Logic Apps.

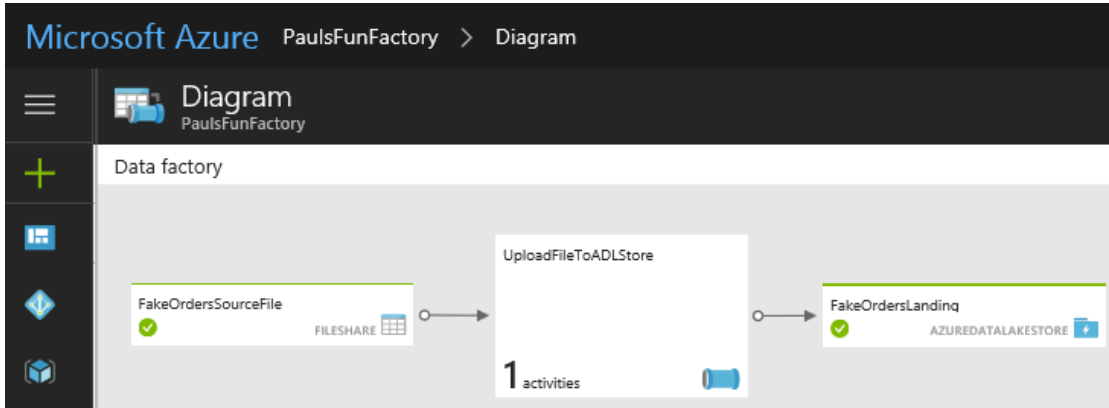
Time Slice Relationships in ADF and Provisioning

```
//Input Dataset
"availability": {
   "frequency": "Day",
   "interval": 1
    }
```

```
//Activity
"scheduler": {
    "frequency": "Day",
    "interval": 1
    }
```

```
//Output Dataset
"availability": {
  "frequency": "Day",
  "interval": 1
   }
```

```
//Pipeline
"start": "2017-05-04T00:00:00Z",
"end": "2018-05-04T00:00:00Z",
"isPaused": true,
"pipelineMode": "Scheduled"
}
```



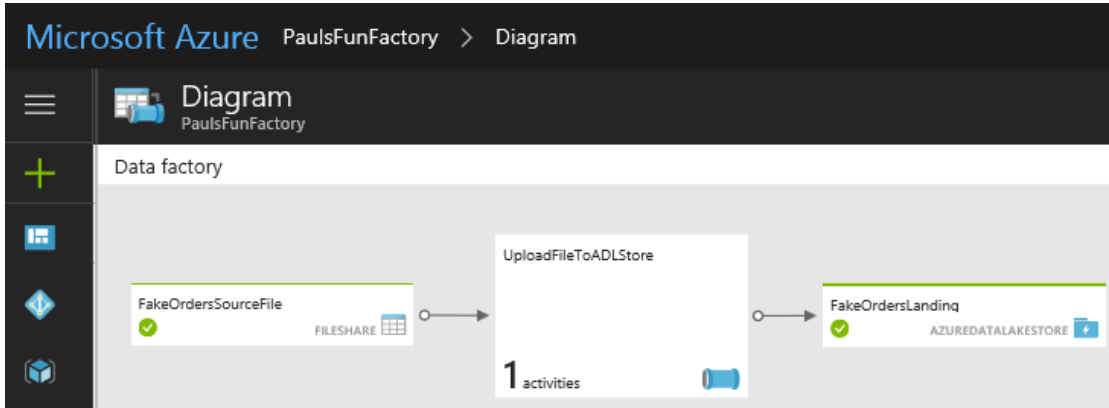
Time Slice Relationships in ADF and Provisioning

```
//Input Dataset
"availability": {
    "frequency": "Day",
    "interval": 1
}
```

```
//Activity
"scheduler": {
    "frequency": "Day",
    "interval": 1
}
```

```
//Output Dataset
"availability": {
    "frequency": "Day",
    "interval": 1
}
```

```
//Pipeline
"start": "2017-05-04T00:00:00Z",
"end": "2018-05-04T00:00:00Z",
"isPaused": true,
"pipelineMode": "Scheduled"
}
```

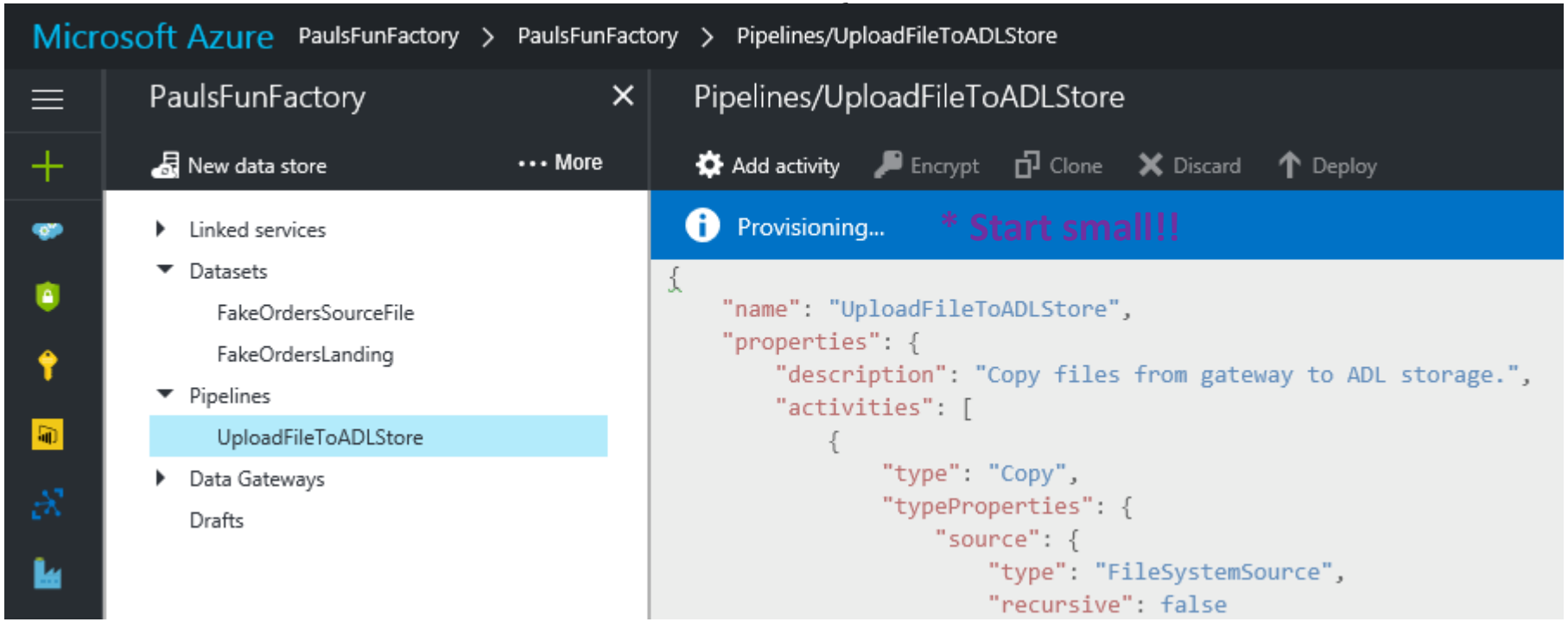


Time Slice Relationships in ADF and Provisioning

```
//Input Dataset
"availability": {
   "frequency": "Day",
   "interval": 1
    }
```

```
//Output Dataset
"availability": {
   "frequency": "Day",
   "interval": 1
    }
```

```
//Activity
```

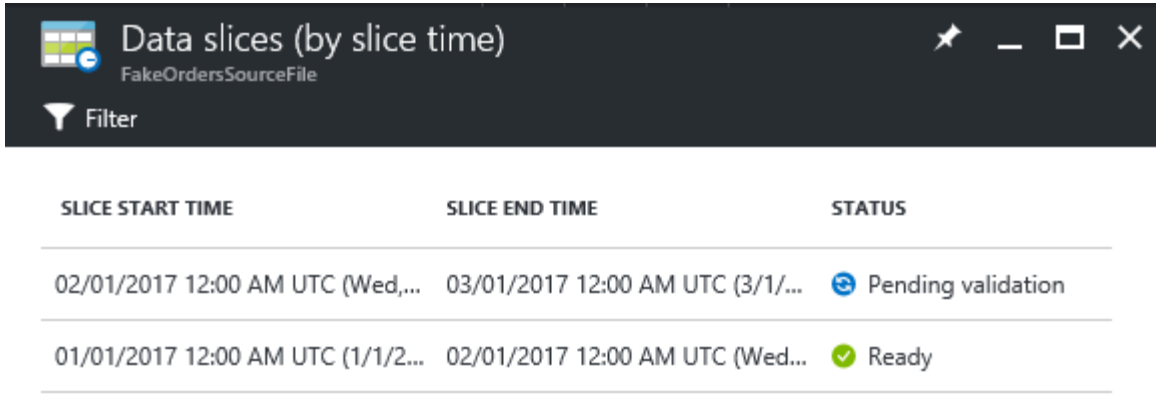


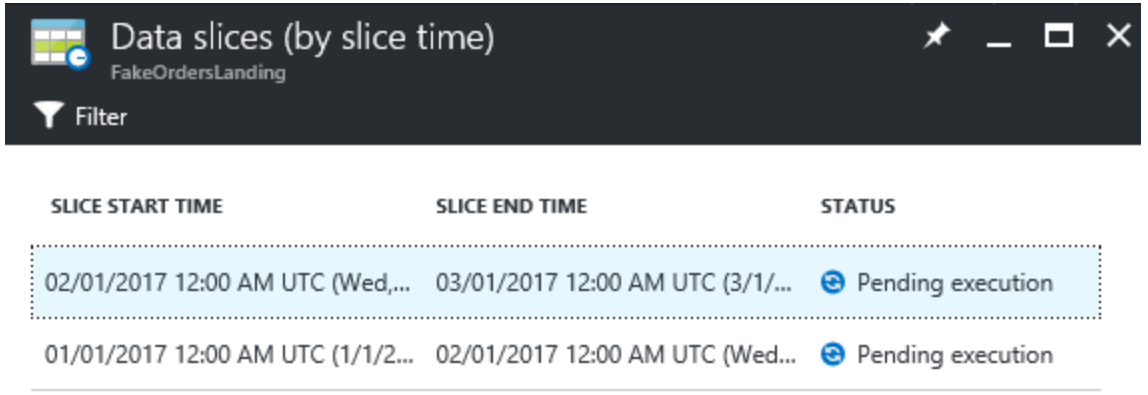
Time Slice Relationships in ADF and Provisioning

```
//Input Dataset
"availability": {
   "frequency": "Day",
   "interval": 1
    }
```

```
//Activity
"scheduler": {
    "frequency": "Day",
    "interval": 1
    }
```

```
//Output Dataset
"availability": {
  "frequency": "Day",
  "interval": 1
   }
```



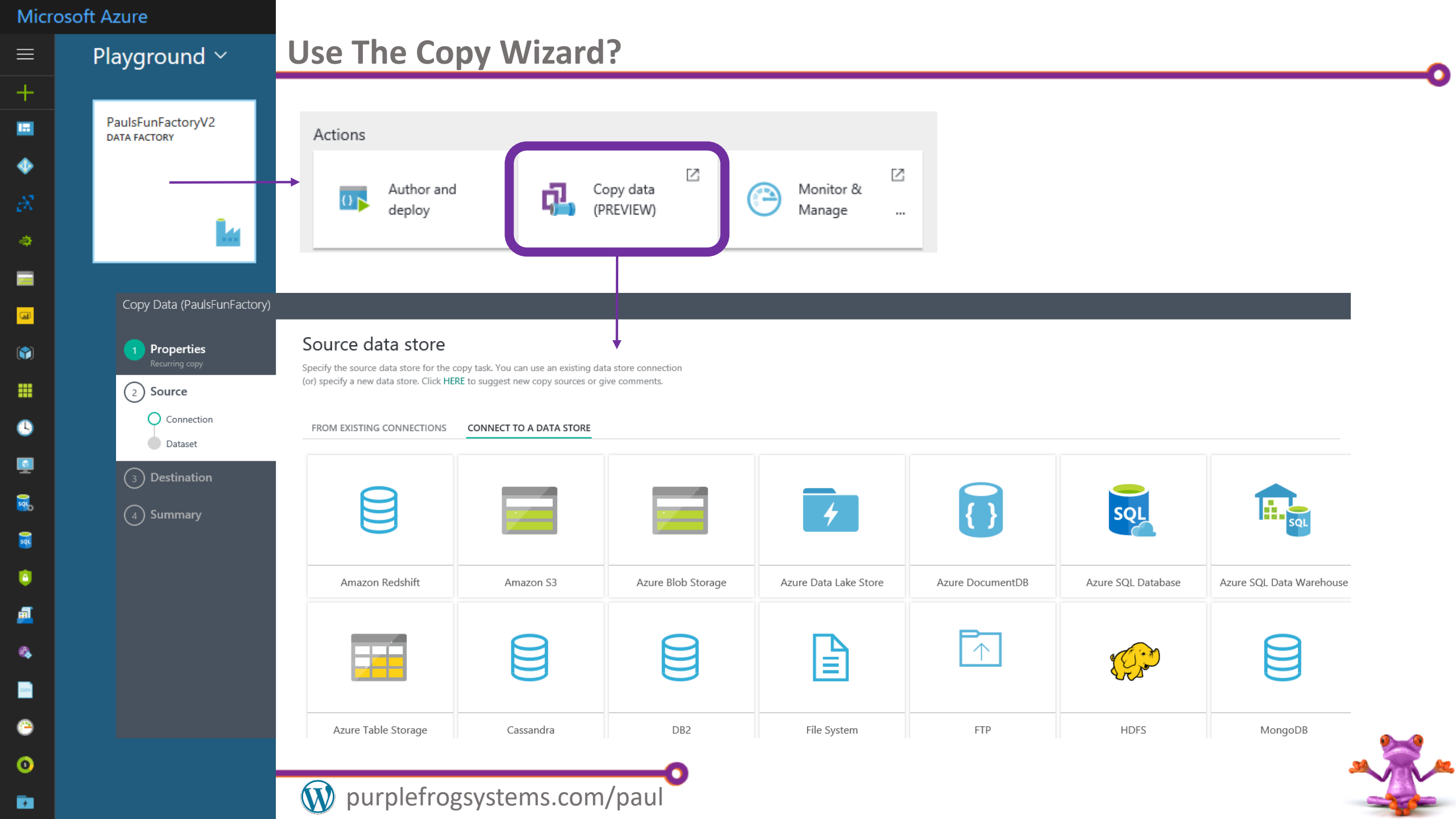


Recap

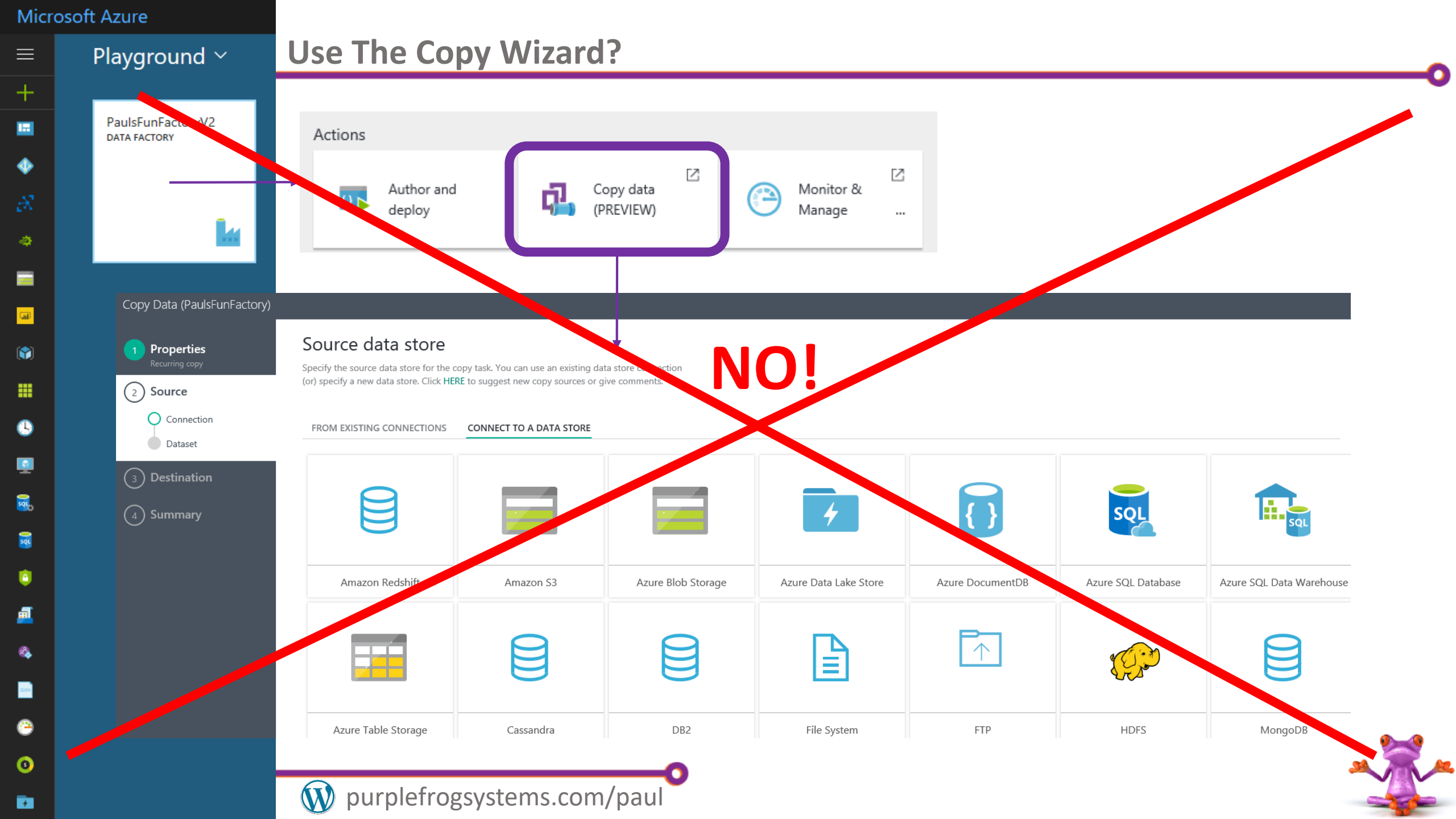
1. What is Azure Data Factory (ADF)? ✓
   - What is an ADF Pipeline? ✓
   - What is an ADF Dataset? ✓
2. What is an ADF Activity? ✓
3. What is an ADF Time Slice? ✓

What do you do next?

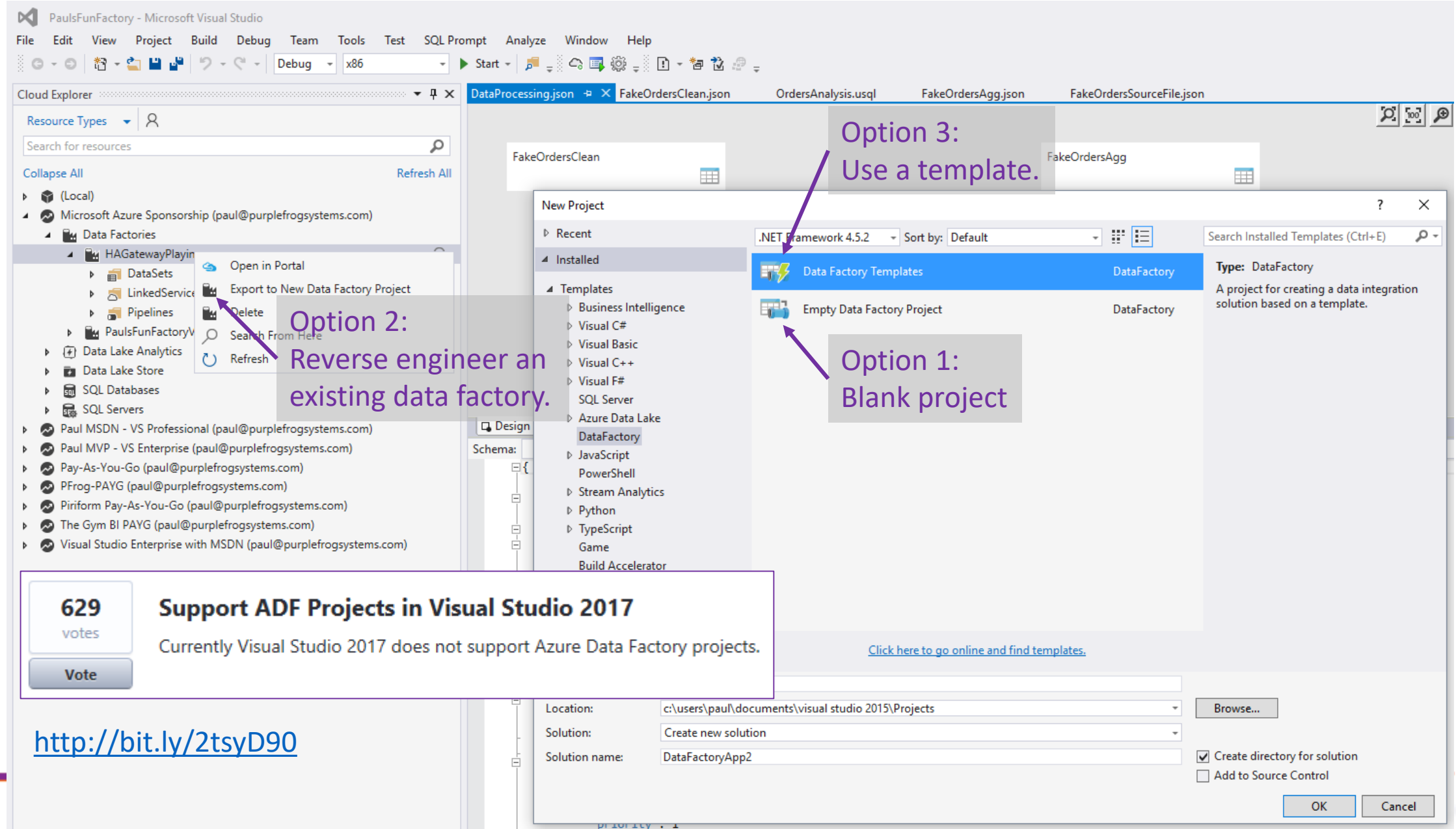
Use The Copy Wizard?



Use The Copy Wizard?



Use Visual Studio 2015



Visual Studio 2015

Visual Studio 2017

629 votes

Vote

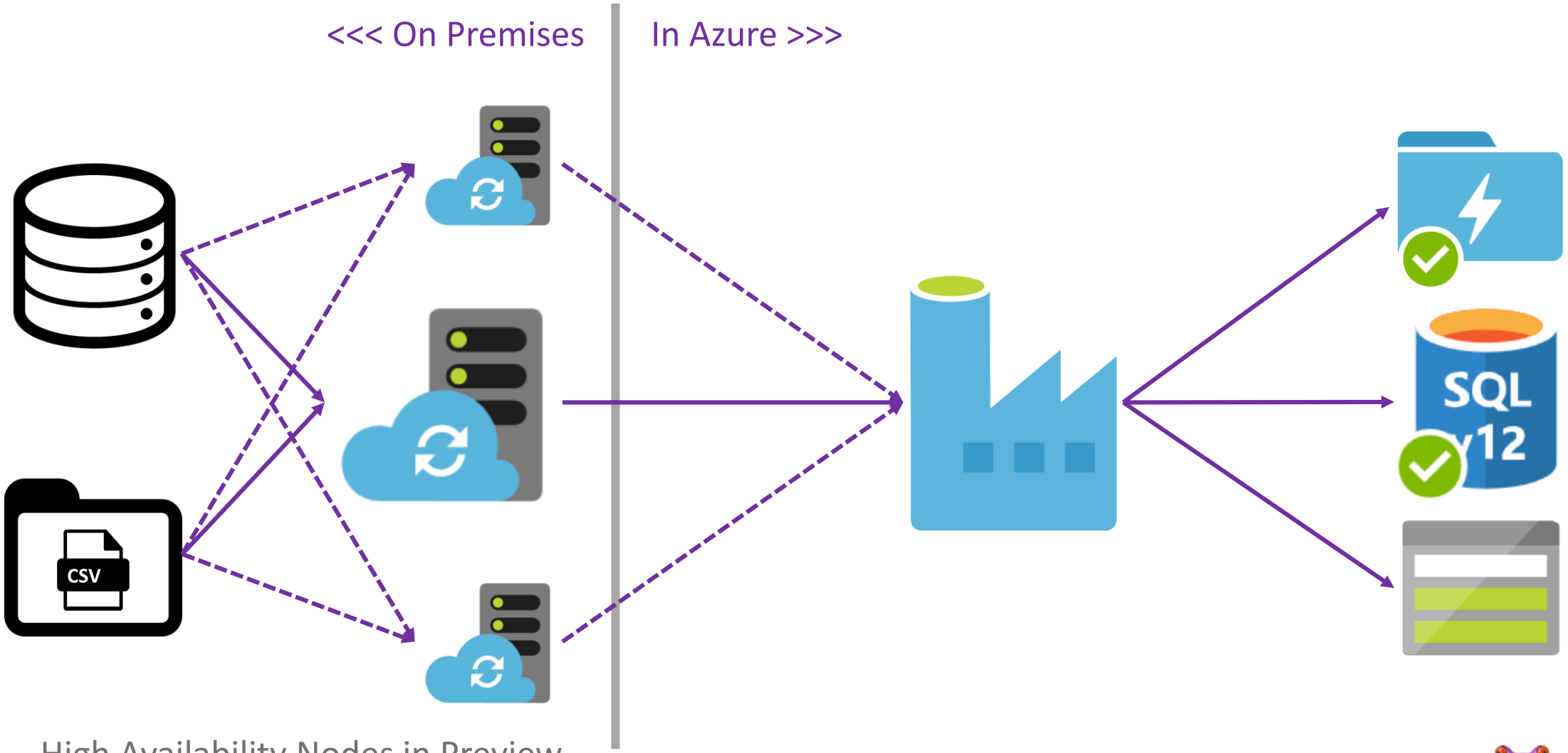
Support ADF Projects in Visual Studio 2017

Currently Visual Studio 2017 does not support Azure Data Factory projects.

http://bit.ly/2tsyD90

The Data Management Gateway

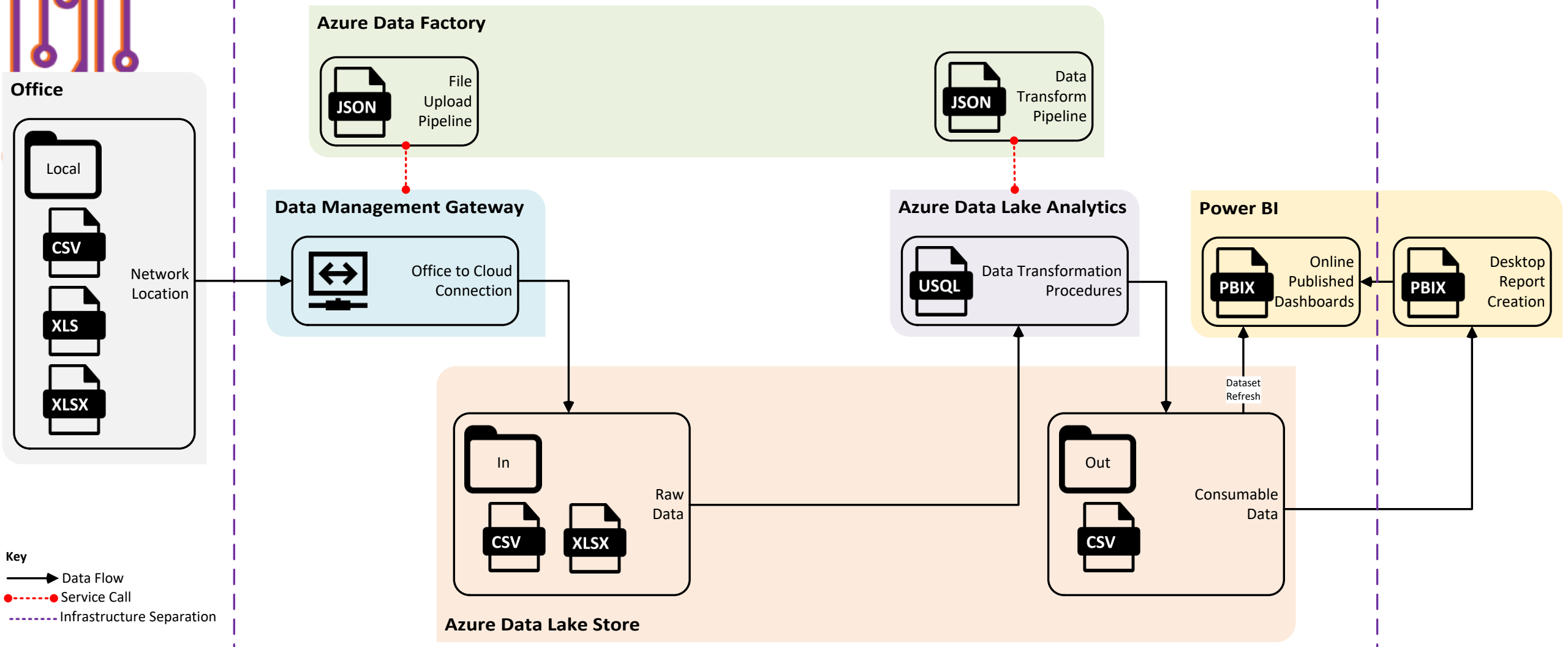
… nothing to do with Power BI.



High Availability Nodes in Preview

Use Case Example One

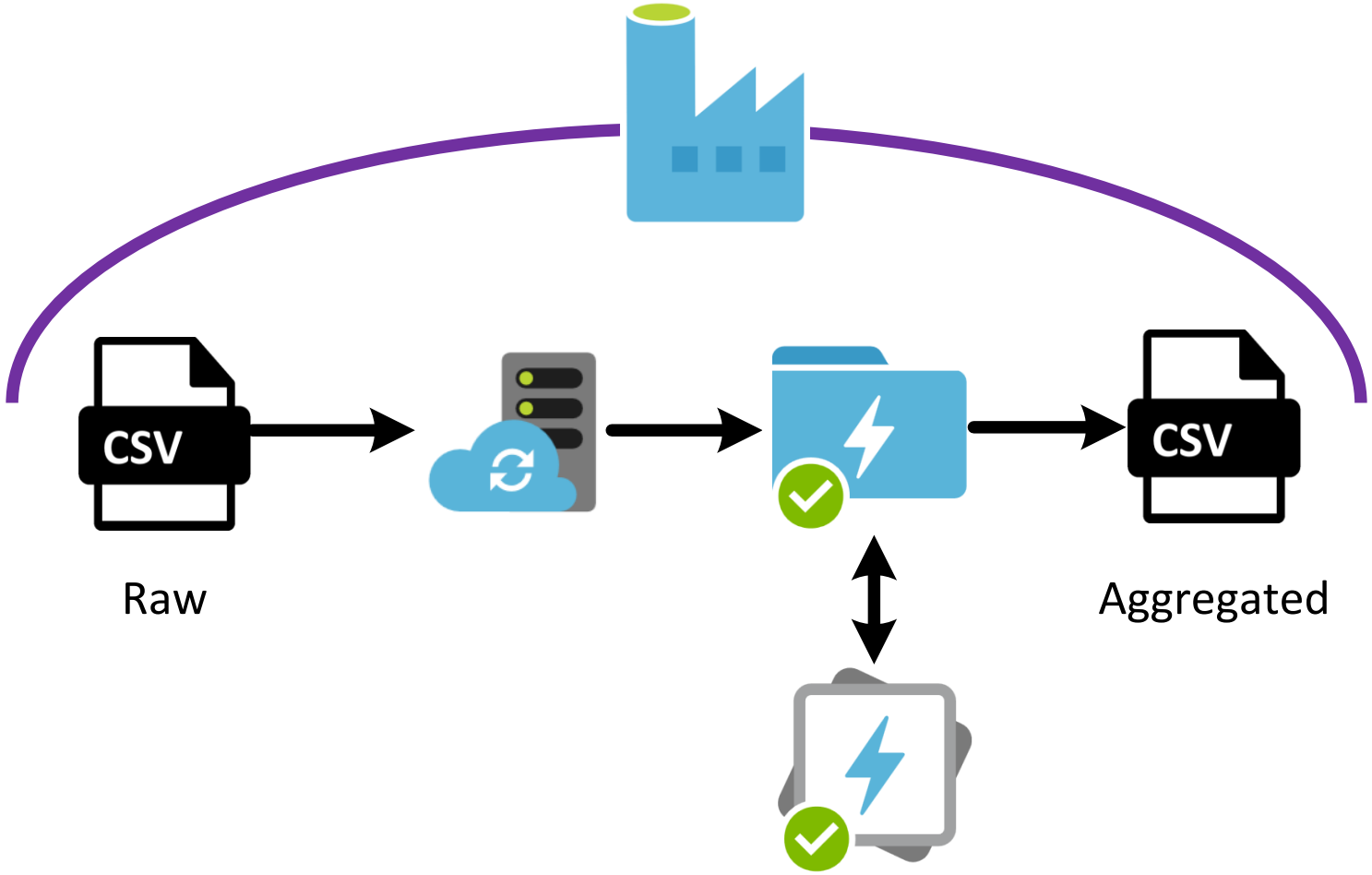
A Purple Frog Story



* ELT – Extract Load Transform

Demo Time

Demo Architecture Version 1



Azure Data Factory – Part 2

with Paul Andrew

Session Agenda

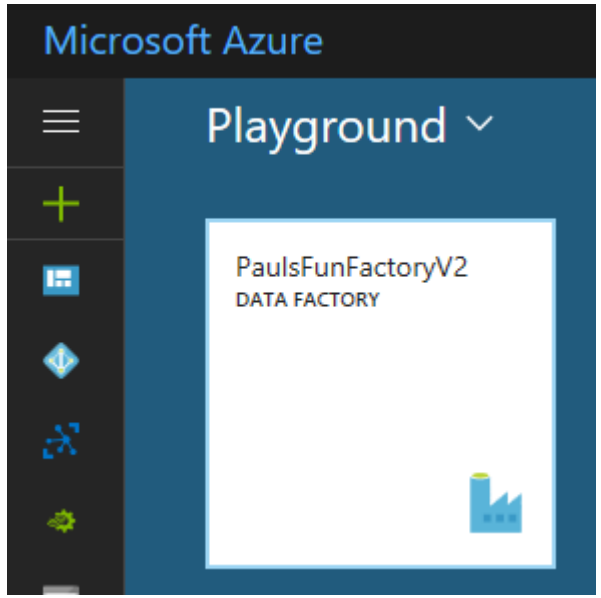
Part 1

1. What is Azure Data Factory (ADF)?
   - What is an ADF Pipeline?
   - What is an ADF Dataset?
2. What is an ADF Activity?
3. What is an ADF Time Slice?
4. Developer tool for ADF.
5. The ADF Data Management Gateway.
6. Building a basic copy data factory.

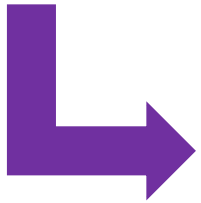
Part 2

7. What is a ADF Custom Activity?
8. Creating a custom activity.
9. Using Visual Studio ADF config files.
10. Monitoring your ADF pipelines.
11. Q&A

What is an ADF Custom Activity?

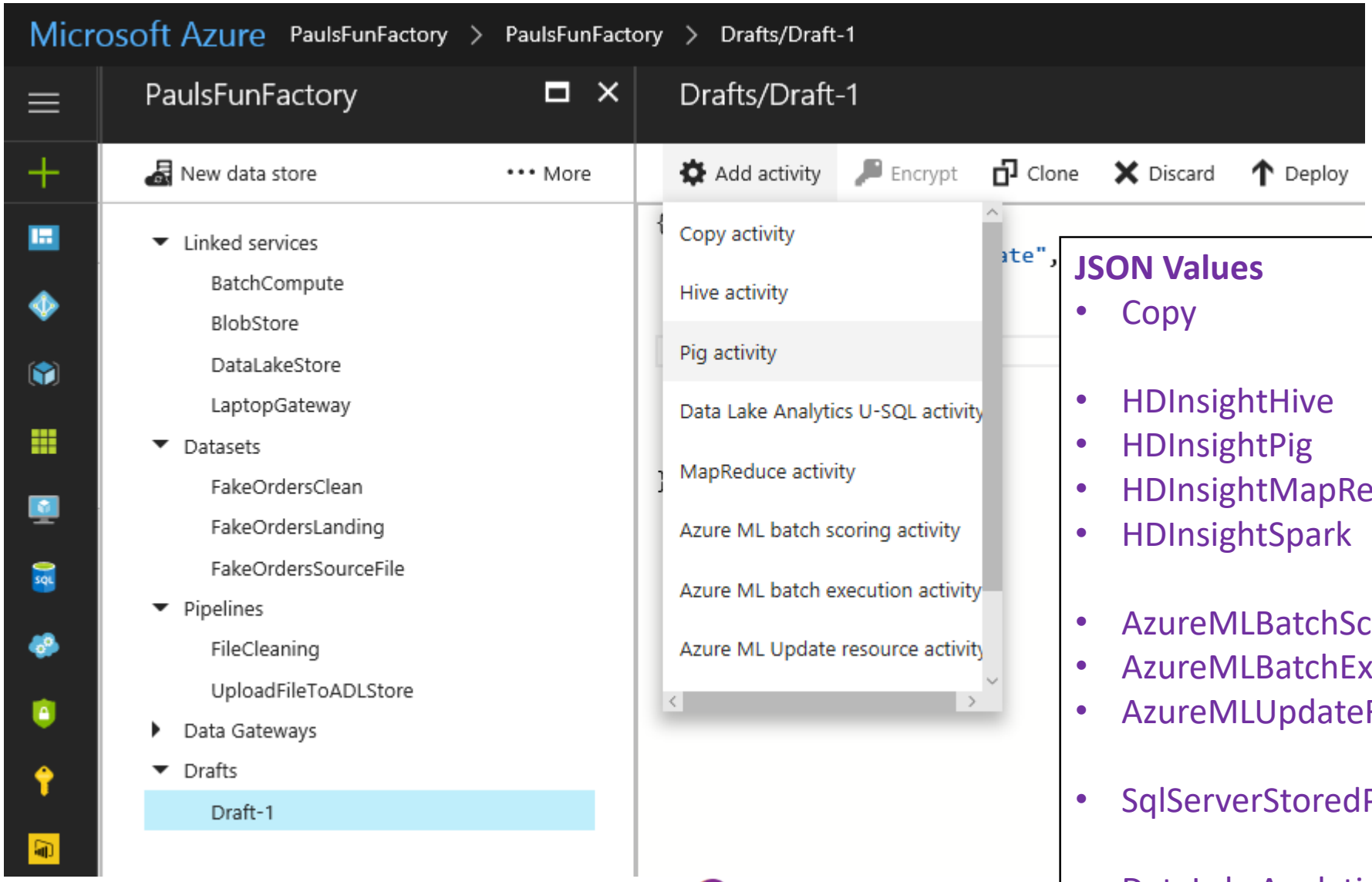


Blog post:
https://www.purplefrogsystems.com/paul/2016/11/creating-azure-data-factory-custom-activities/



```
{
    "$schema": "http://datafactories.schema
    "name": "CustomActivity01",
    "properties": {
        "description": "Going Beyond!",
        "activities": [
            {
                "name": "The DLL Eater",
                "type": "DotNetActivity",
                "inputs": [
                    {
```

How to Create an ADF Custom Activity?



JSON Values

- Copy

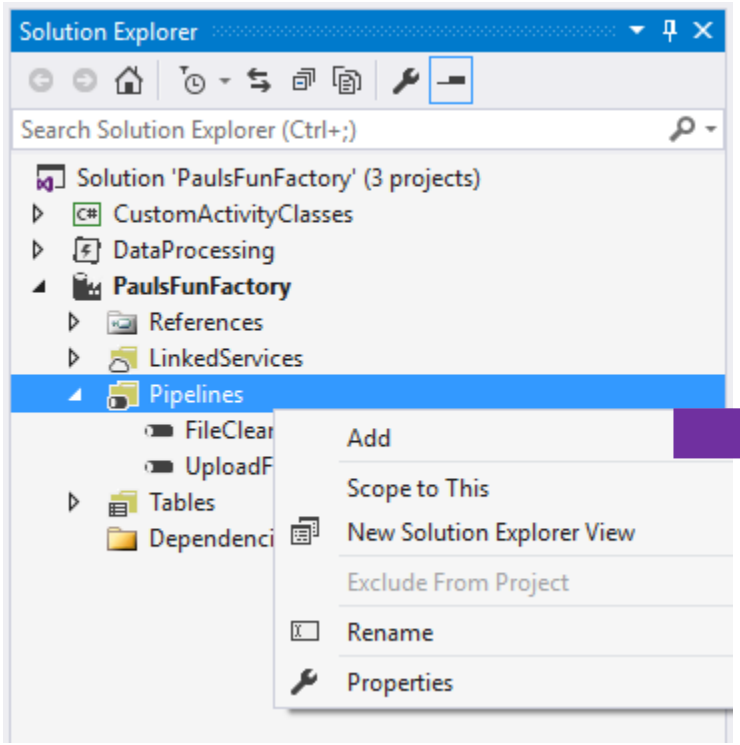
- HDInsightHive
- HDInsightPig
- HDInsightMapReduce
- HDInsightSpark

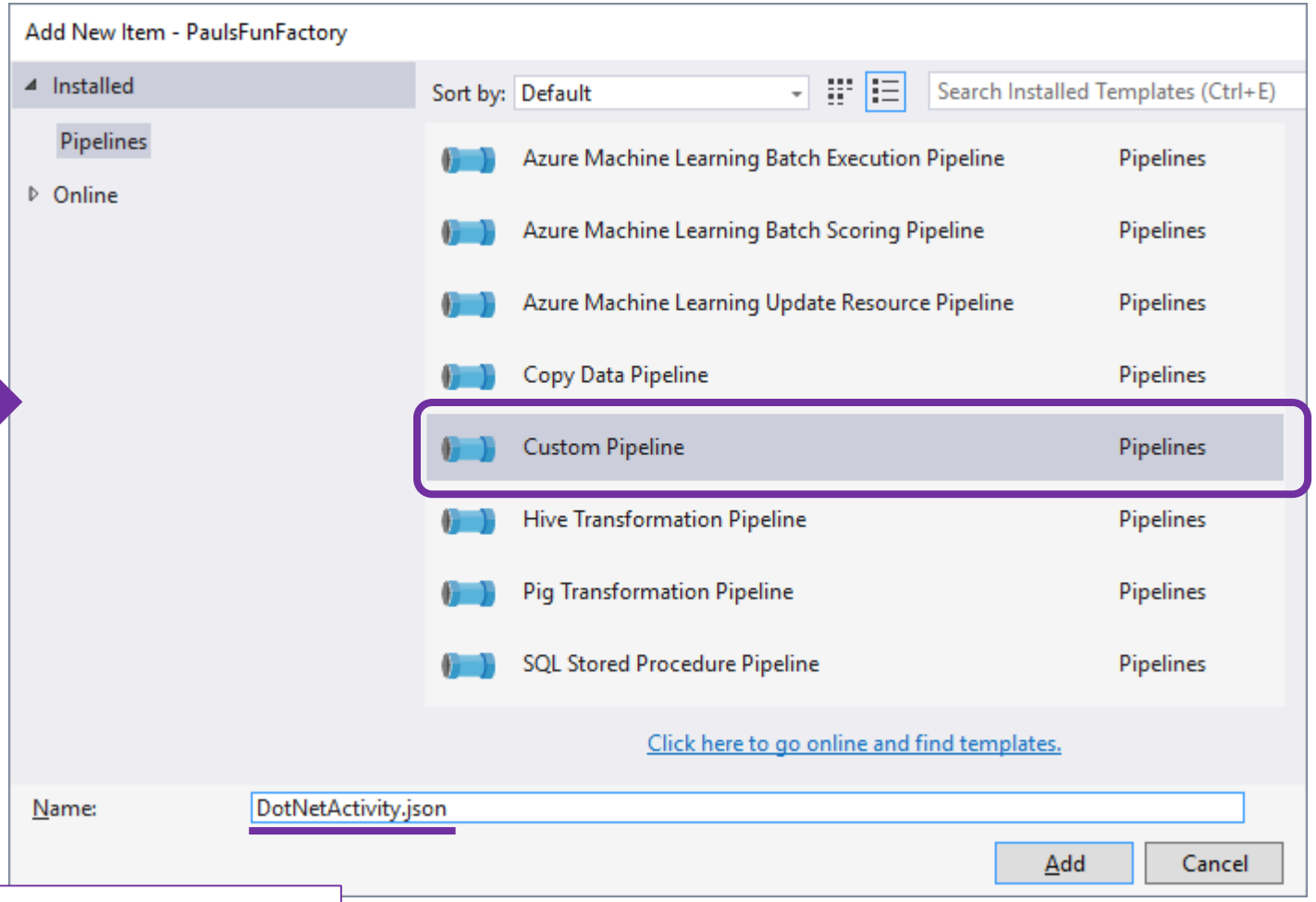
- AzureMLBatchScoring
- AzureMLBatchExecution
- AzureMLUpdateResource

- SqlServerStoredProcedure

- DataLakeAnalyticsU-SQL

How to Create an ADF Custom Activity?





Visual Studio 2015

Visual Studio 2017

http://bit.ly/2tsyD90

629 votes

Vote

Support ADF Projects in Visual Studio 2017

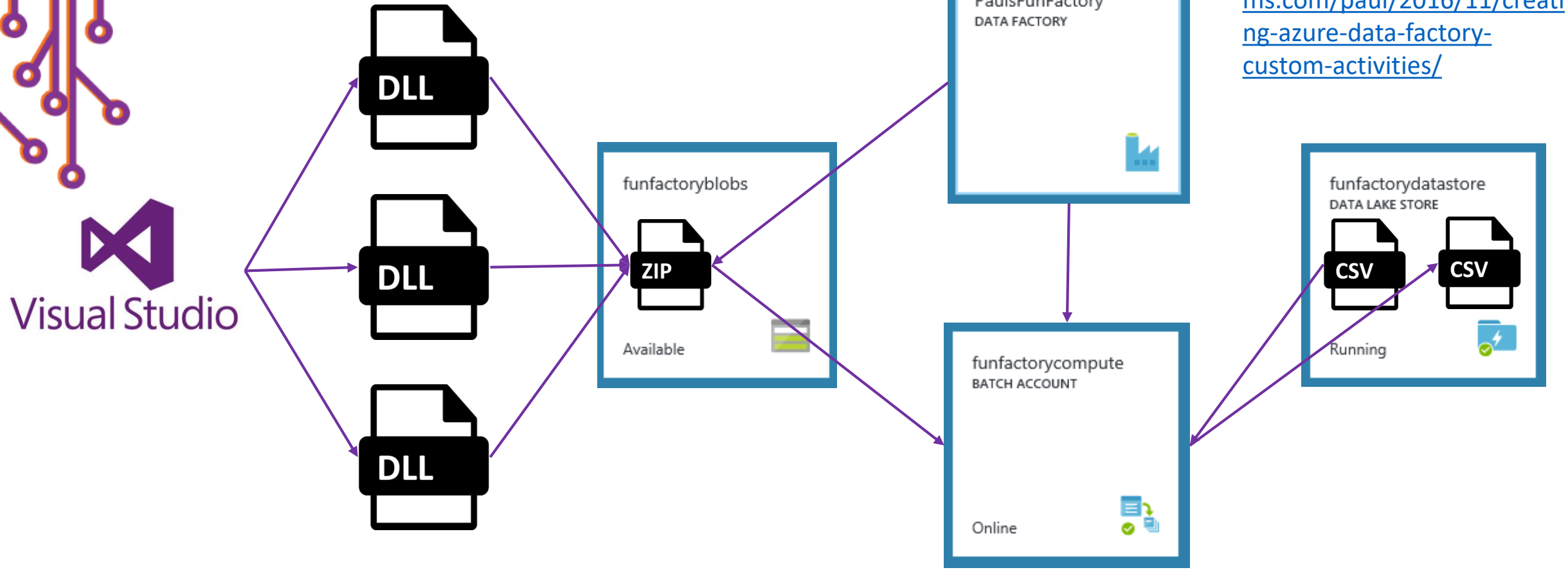
Currently Visual Studio 2017 does not support Azure Data Factory projects.

ADF Custom Activity Properties

```
{
    "$schema": "http://datafactories.schema.management.azure.com/schemas/2015-09-01/Microsoft.DataFactory.Pipeline.json",
    "name": "CustomActivity01",
        "properties": {
        "description": "Going Beyond!",
        "activities": [
            {
                "name": "The DLL Eater",
                "type": "DotNetActivity",
                "linkedServiceName": "BatchServiceCompute",
                "inputs": [
                    {
                        "name": "File1ADFStore"
                    }
                ],
                "outputs": [
                    {
                        "name": "File1Cleaned"
                    }
                ],
                "typeProperties": {
                    "assemblyName": "The compiled class library DLL name.",
                    "entryPoint": "The namespace and class to instantiate.",
                    "packageLinkedService": "The blob storage account where the zipped up DLLs live.",
                    "packageFile": "The blob storage container and zip file name.",
                    "extendedProperties": {
                        “TimeSliceStart": "$$Text.Format('{0:yyyyMMdd}',SliceStart)"}
                }
```

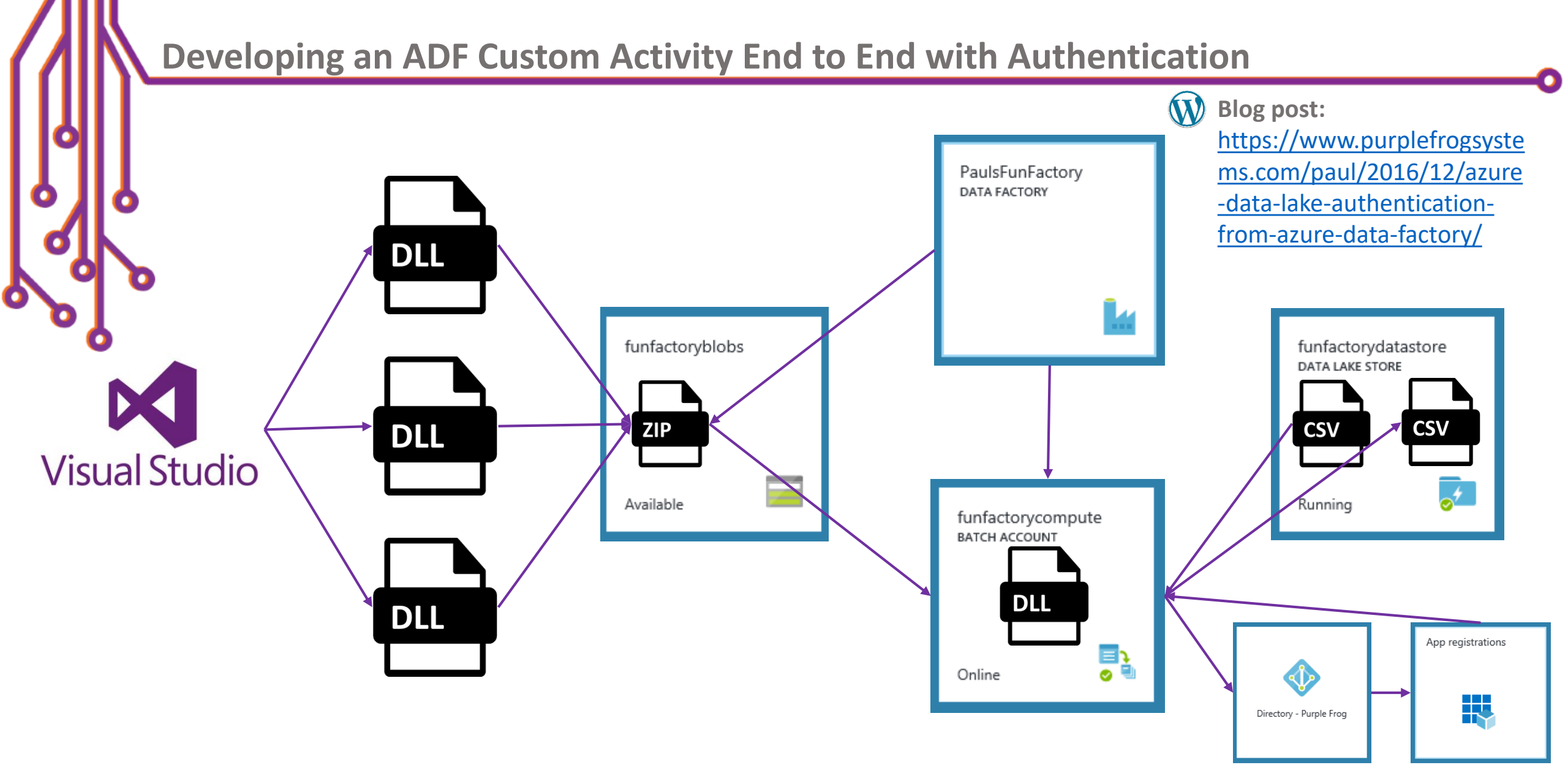
Developing an ADF Custom Activity End to End

Blog post:
https://www.purplefrogsystems.com/paul/2016/11/creating-azure-data-factory-custom-activities/

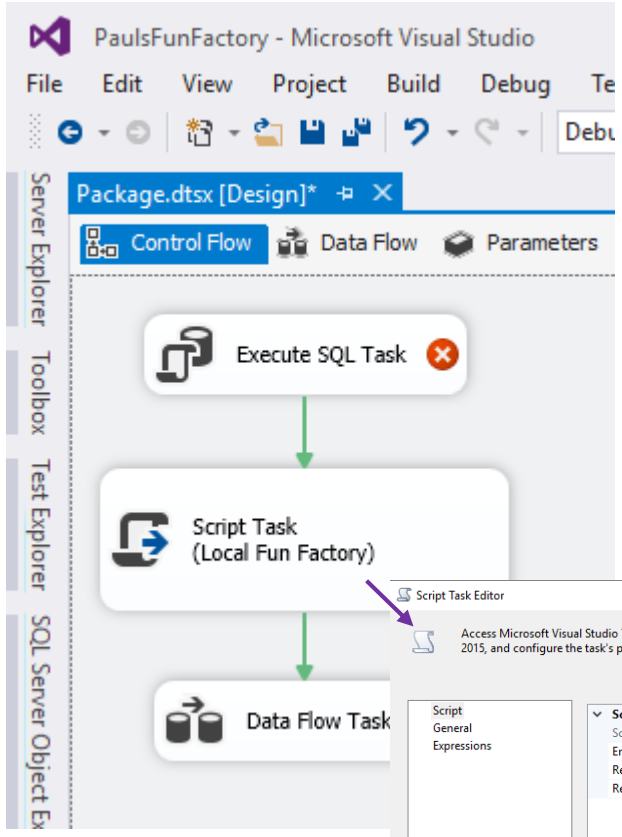


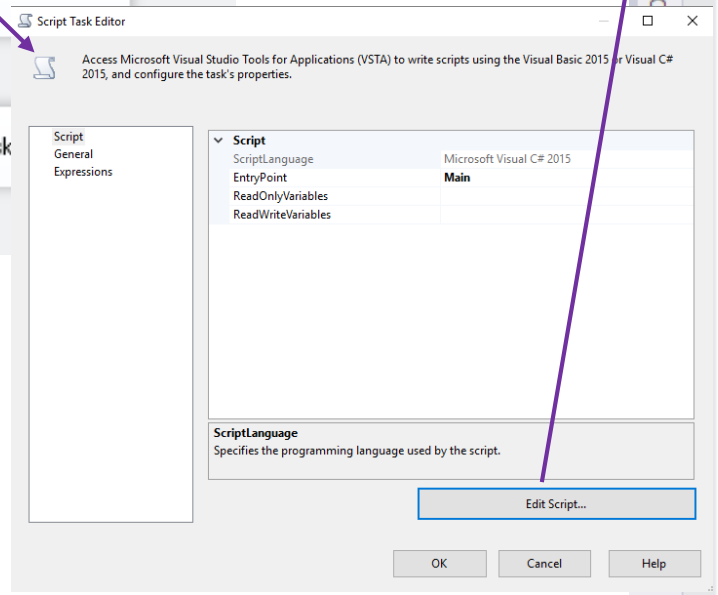
Developing an ADF Custom Activity End to End with Authentication

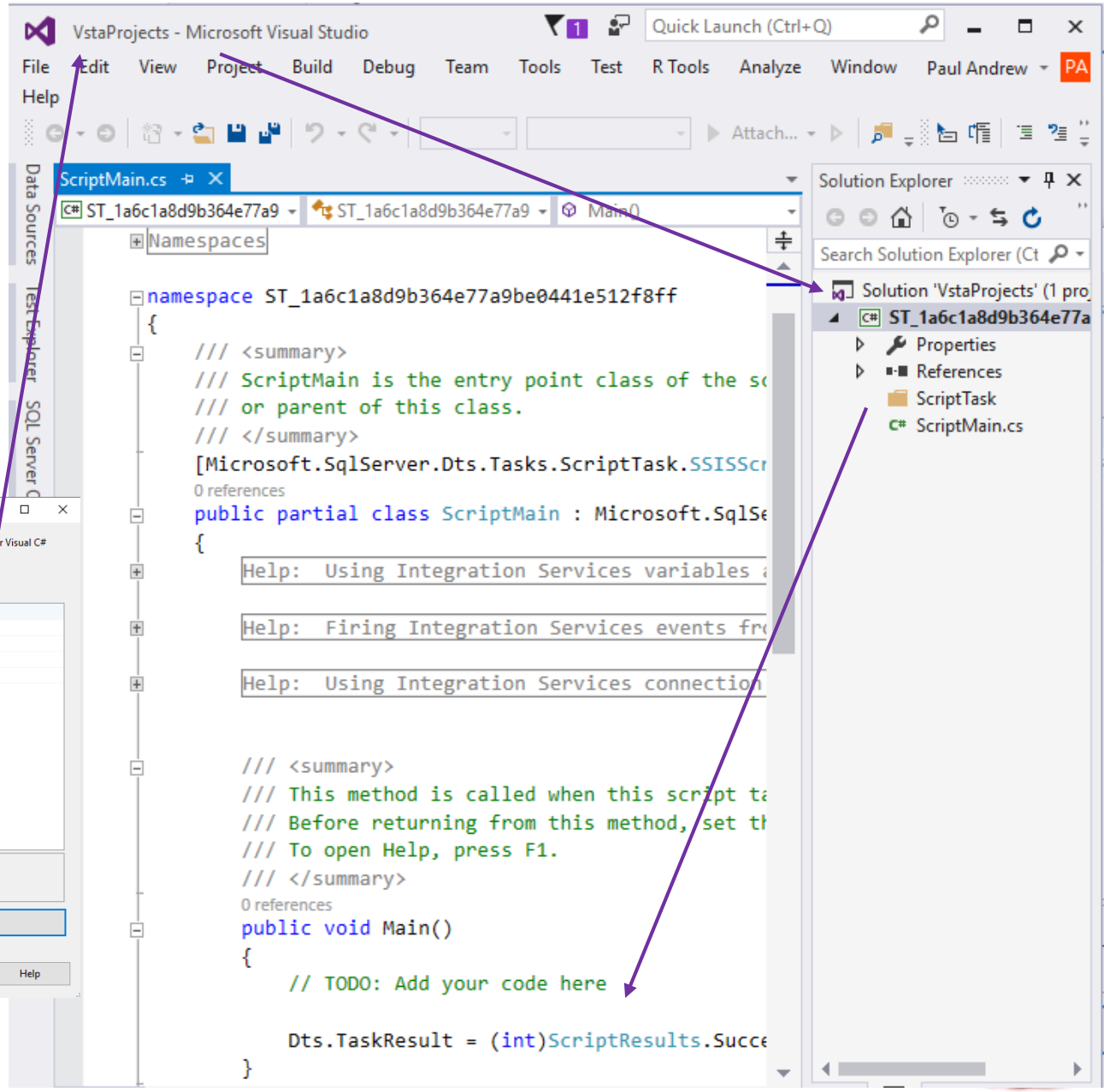
Blog post:
https://www.purplefrogsystems.com/paul/2016/12/azure-data-lake-authentication-from-azure-data-factory/



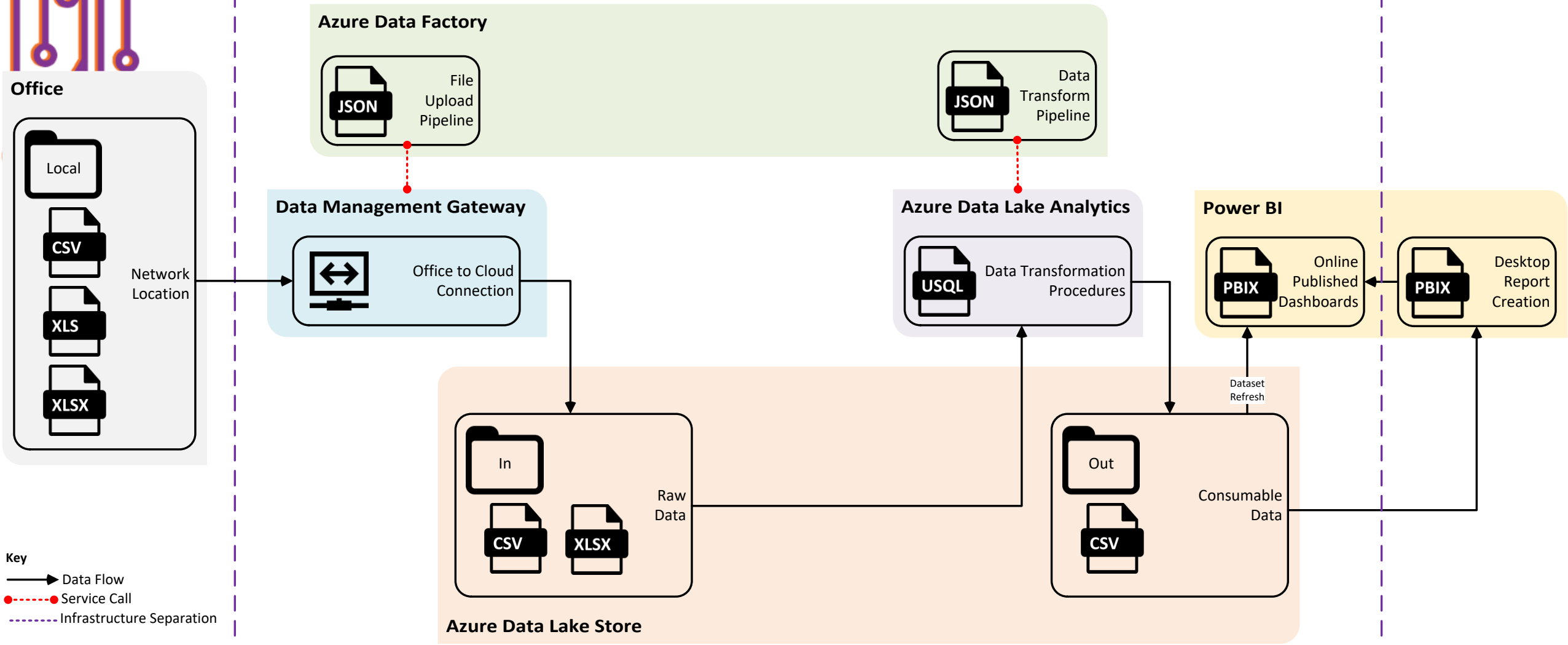
ADF Custom Activity vs SSIS Control Flow Script Task





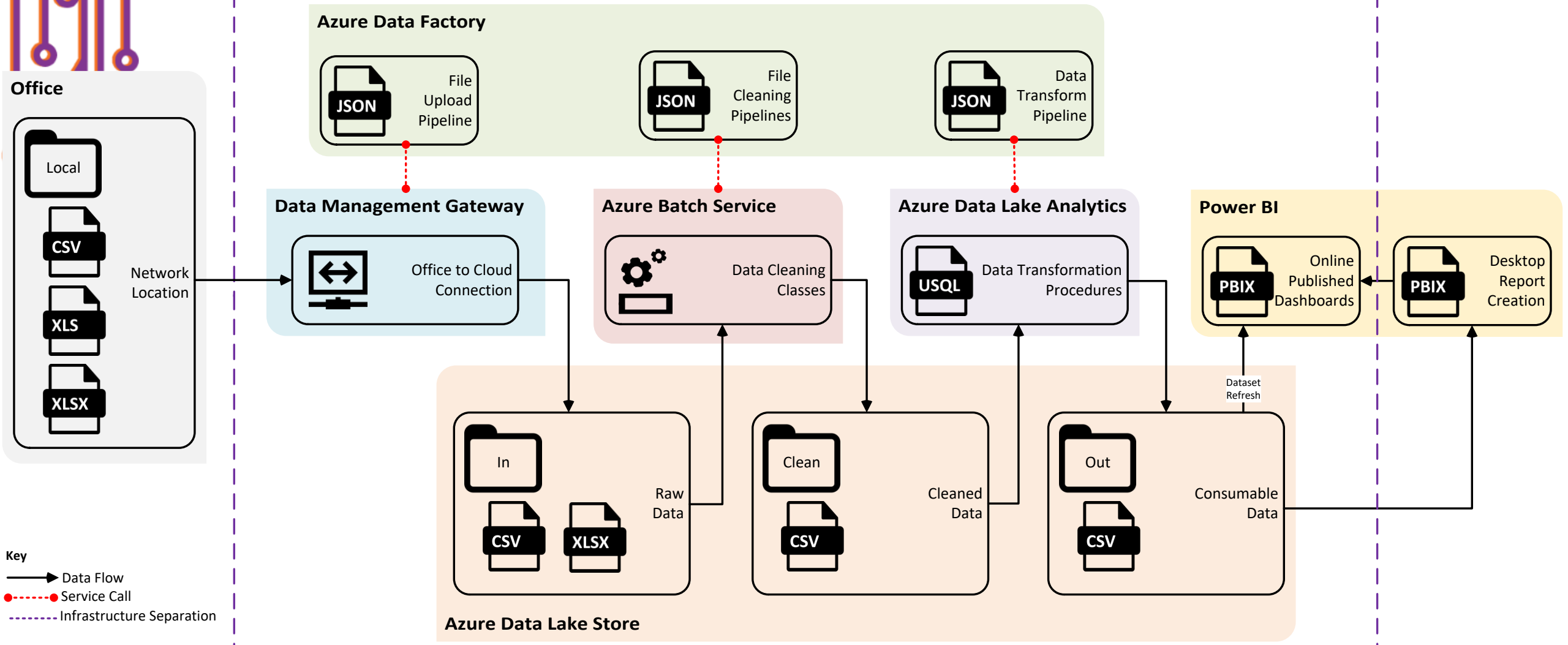


Why do we need an ADF Custom Activity? Example One (Before)



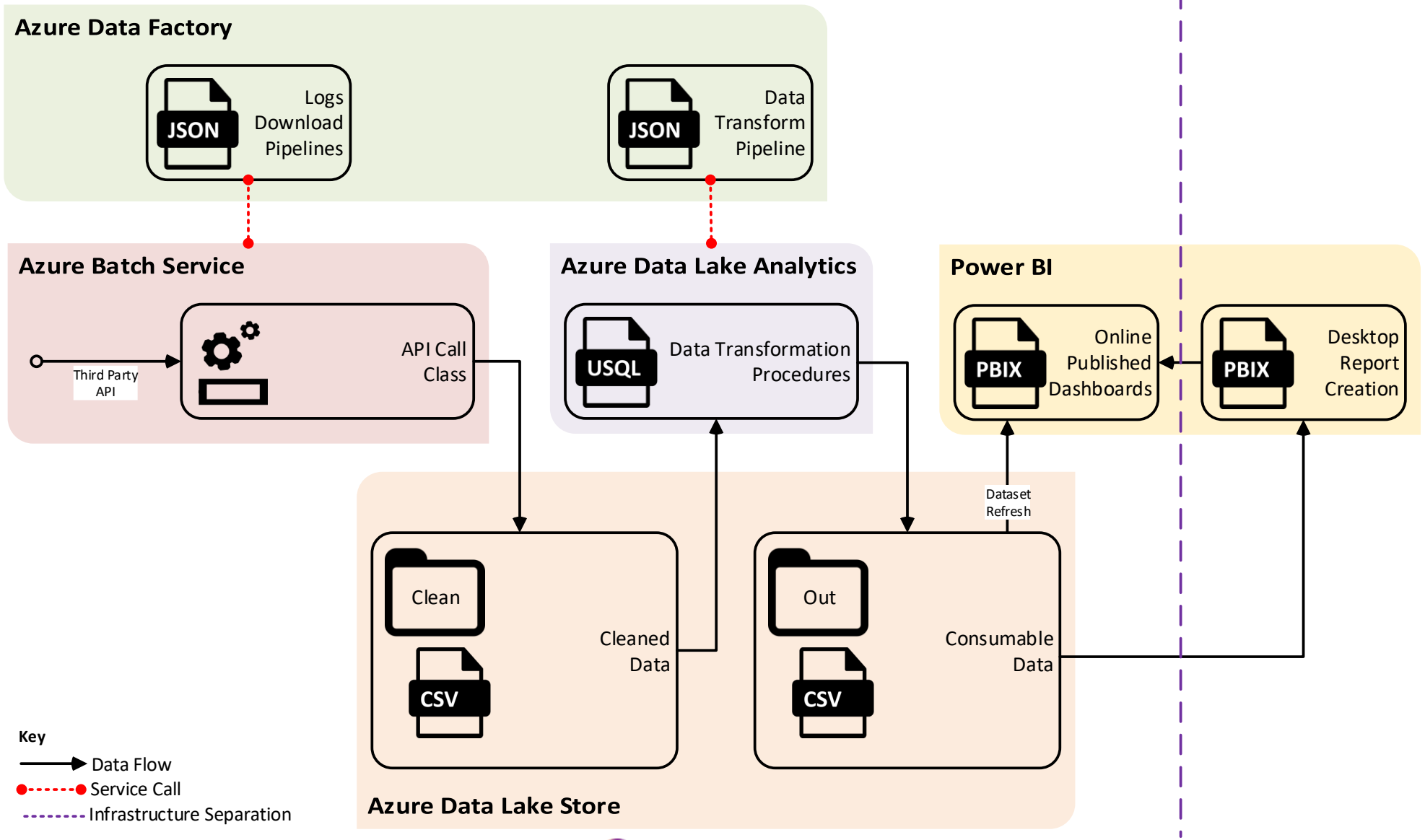
* ELT – Extract Load Transform

Why do we need an ADF Custom Activity? Example One (After)



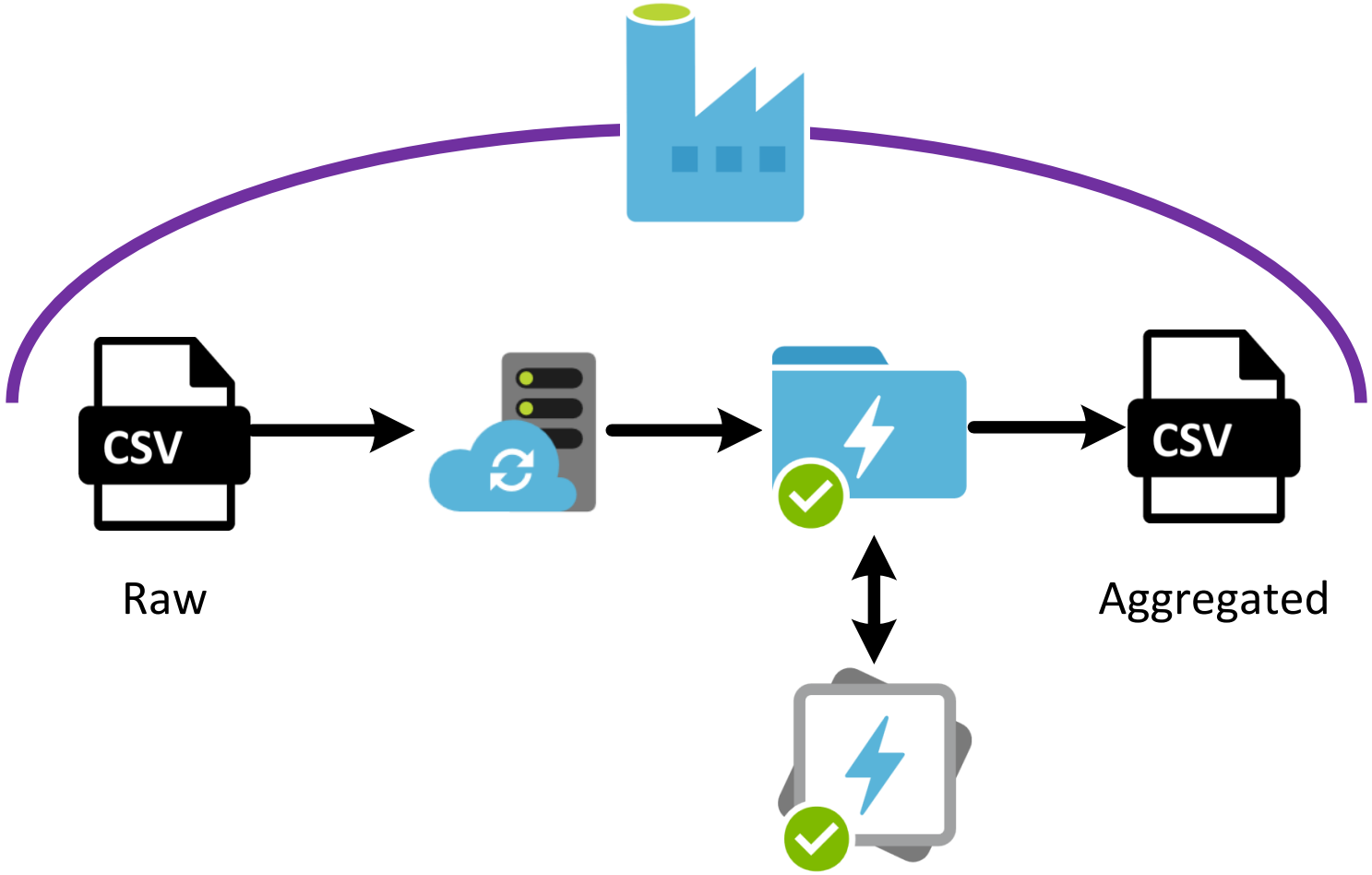
* ECLT – Extract Clean Load Transform

Why do we need an ADF Custom Activity? Example Two

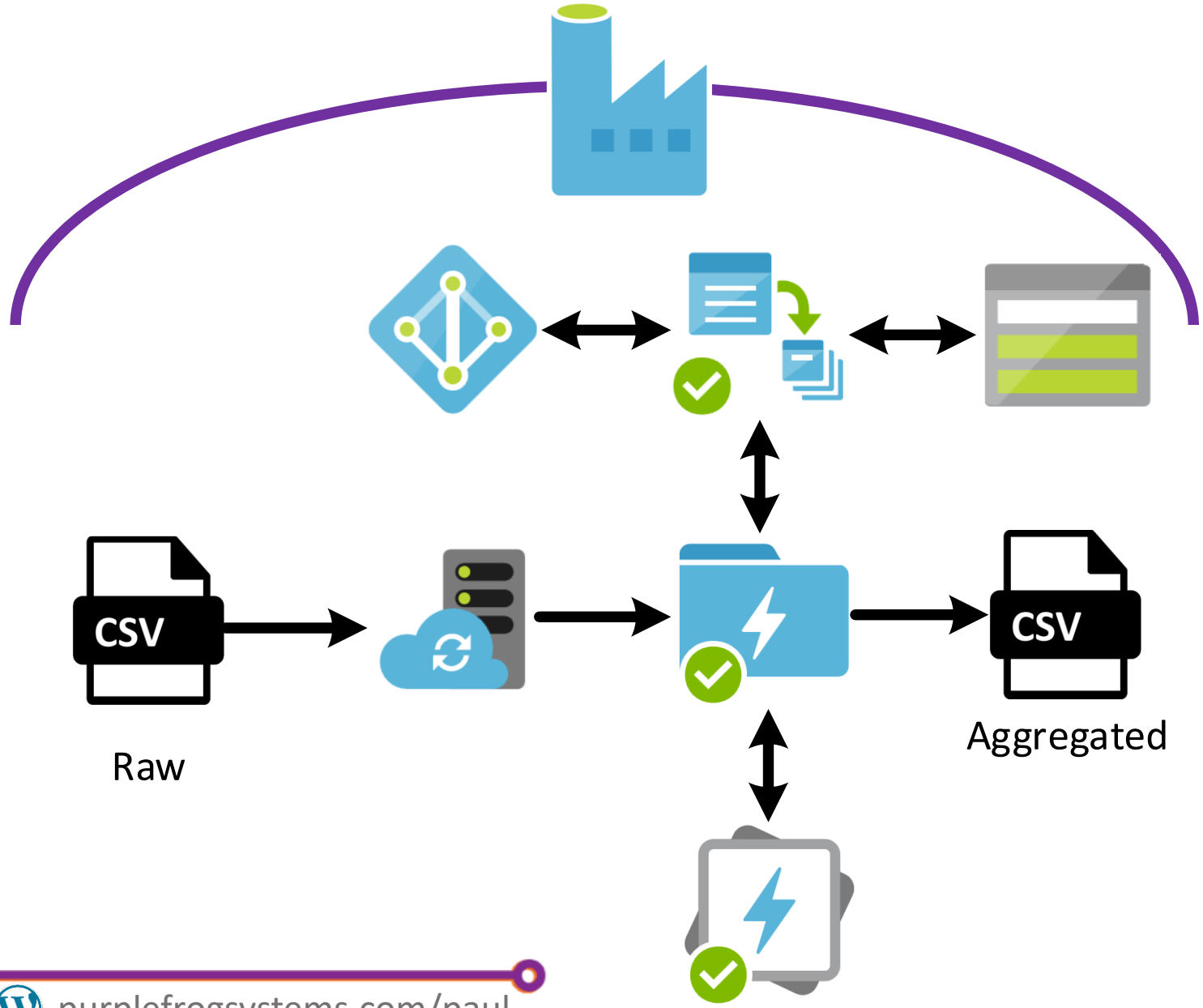


Demo Time

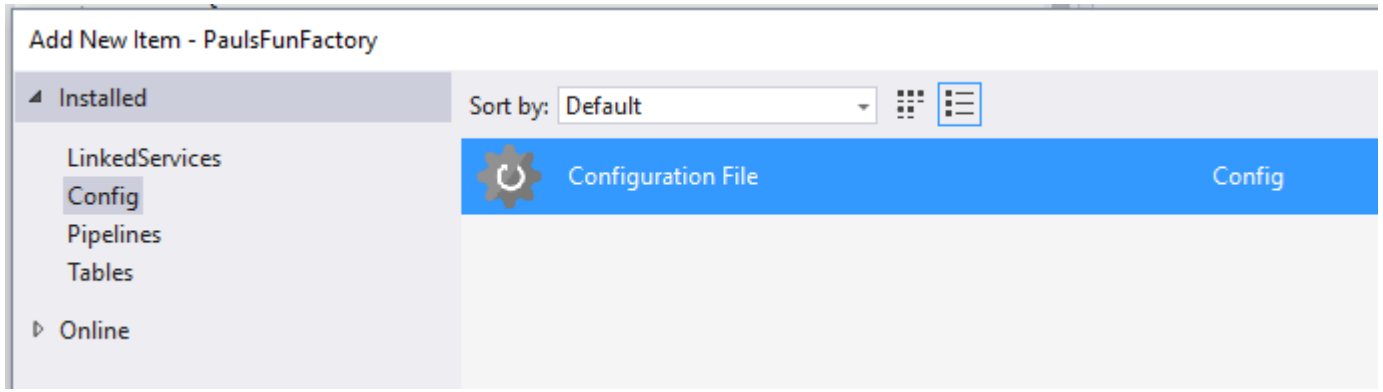
Demo Architecture Version 1



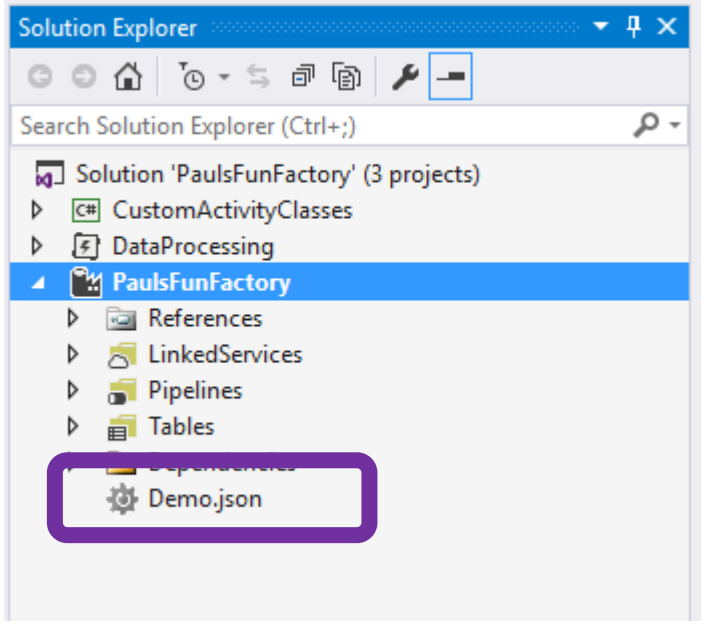
Demo Architecture Version 2



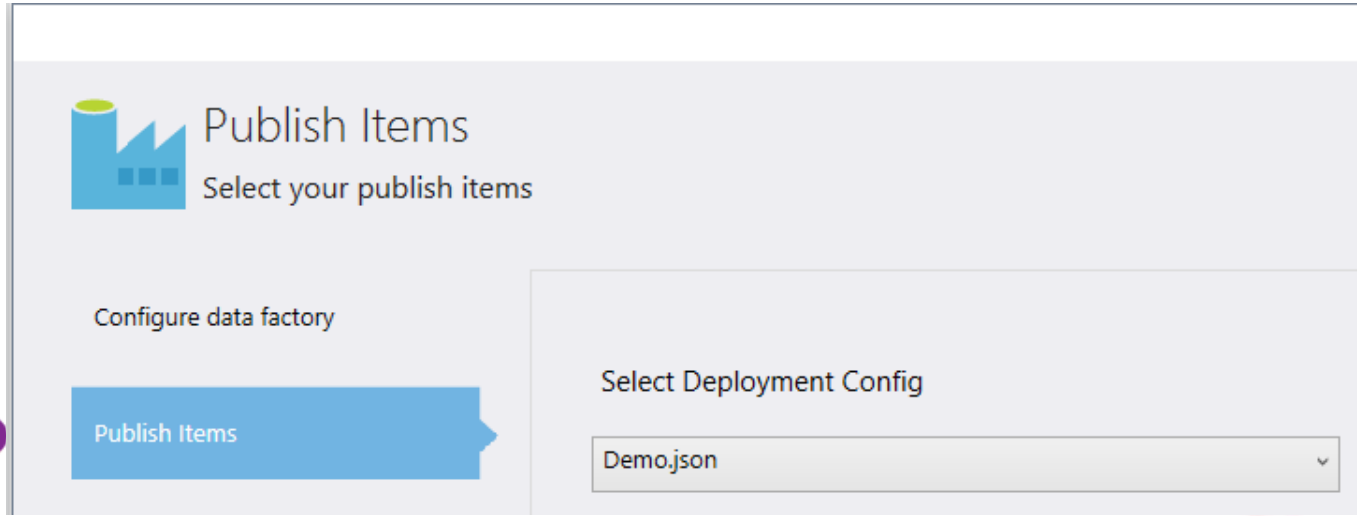
Visual Studio Configuration Files for ADF



Blog post: https://www.purplefrogsystems.com/paul/2017/01/using-azure-data-factory-configuration-files/




```
{
    "$schema": "",
    "<Table/Pipeline/LinkedService name>": [
        {
            "name": "$.JPath", //i.e $.properties.connectionString
            "value": "<value>"
        }
    ]
}
```



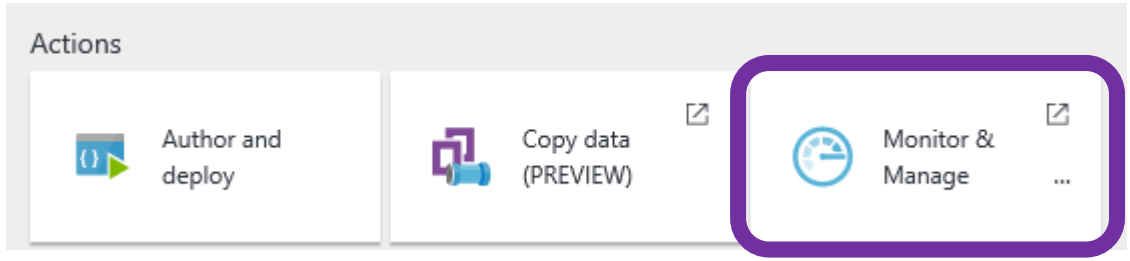
Visual Studio Configuration Files for ADF

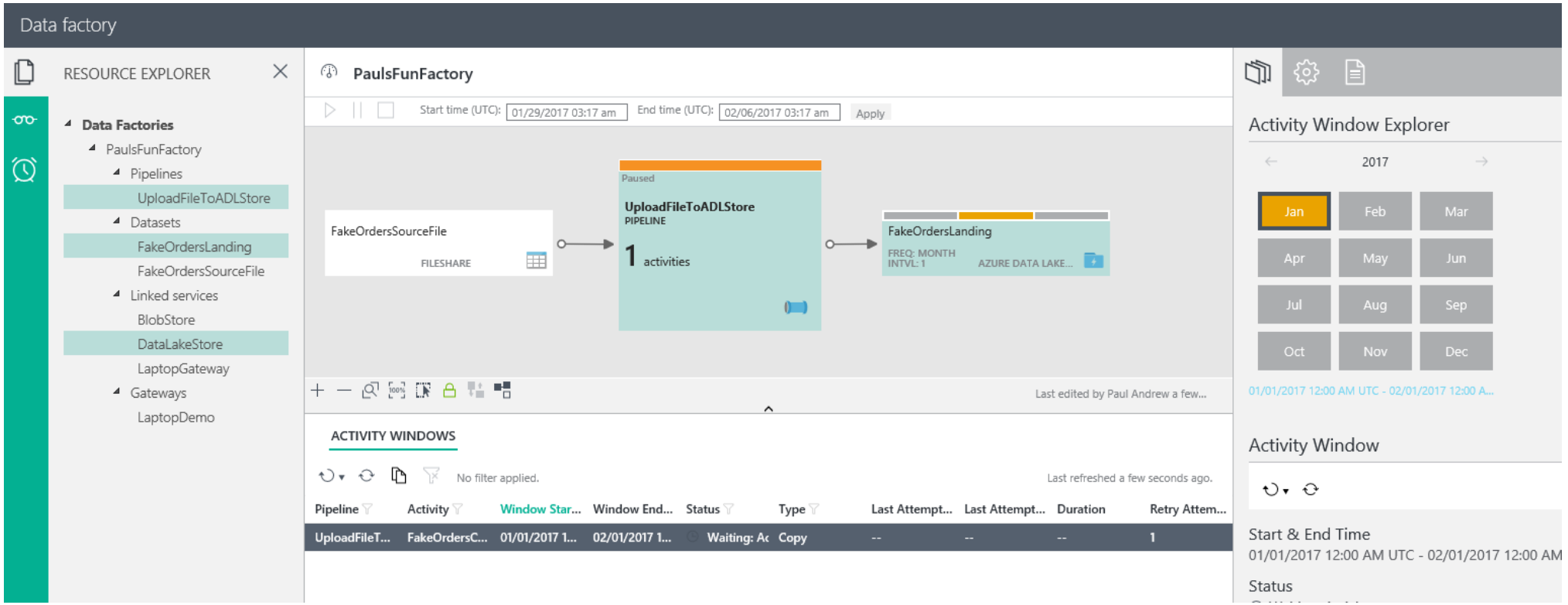
```
//Linked Service:
{
"name": "DataLakeStore",
        "properties": {
                "type": "AzureDataLakeStore",
                "typeProperties": {
                        "accountName": "<get from config file>",
                        "resourceGroupName": "<get from config file>"
                        }
                }
}

//Config:
{
"DataLakeStore": [
        {
        "name": "$.properties.typeProperties.accountName",
        "value": "ccleanerdatalake01"
        },
        {
        "name": "$.properties.typeProperties.resourceGroupName",
        "value": "CCleaner"
        }
        ]
}
```

Blog post:
https://www.purplefrogsystems.com/paul/2017/01/using-azure-data-factory-configuration-files/

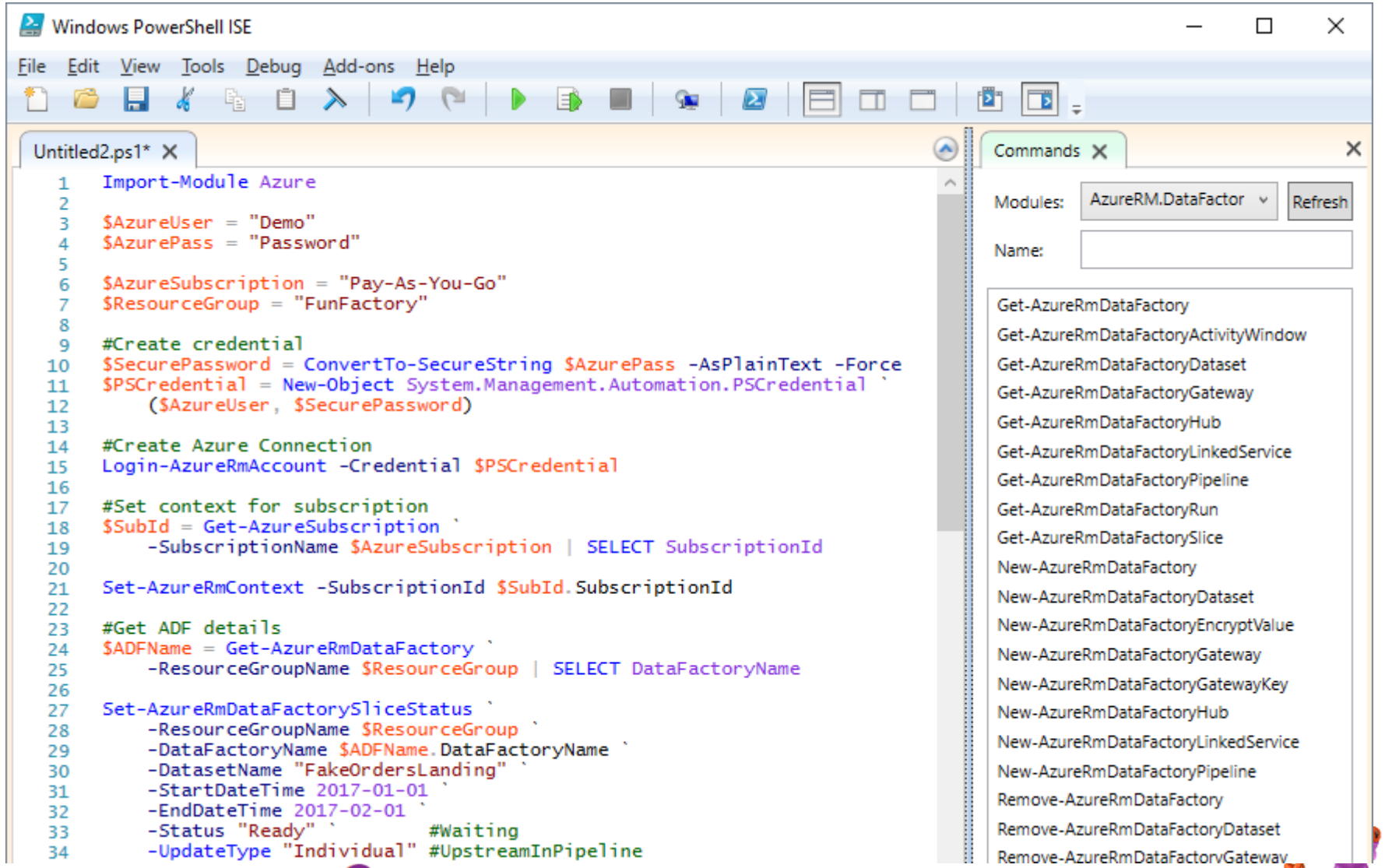
Monitoring, Management and Alerting





PowerShell

https://docs.microsoft.com/en-us/powershell/resourcemanager/azurerm.datafactories/v2.5.0/azurerm.datafactories



[Remove-AzureRmDataFactory]

[Set-AzureRmDataFactorySliceStatus]

[Set-AzureRmDataFactoryGateway]

[Get-AzureRmDataFactoryActivityWindow]

[Get-AzureRmDataFactorySlice]

[Get-AzureRmDataFactoryRun]

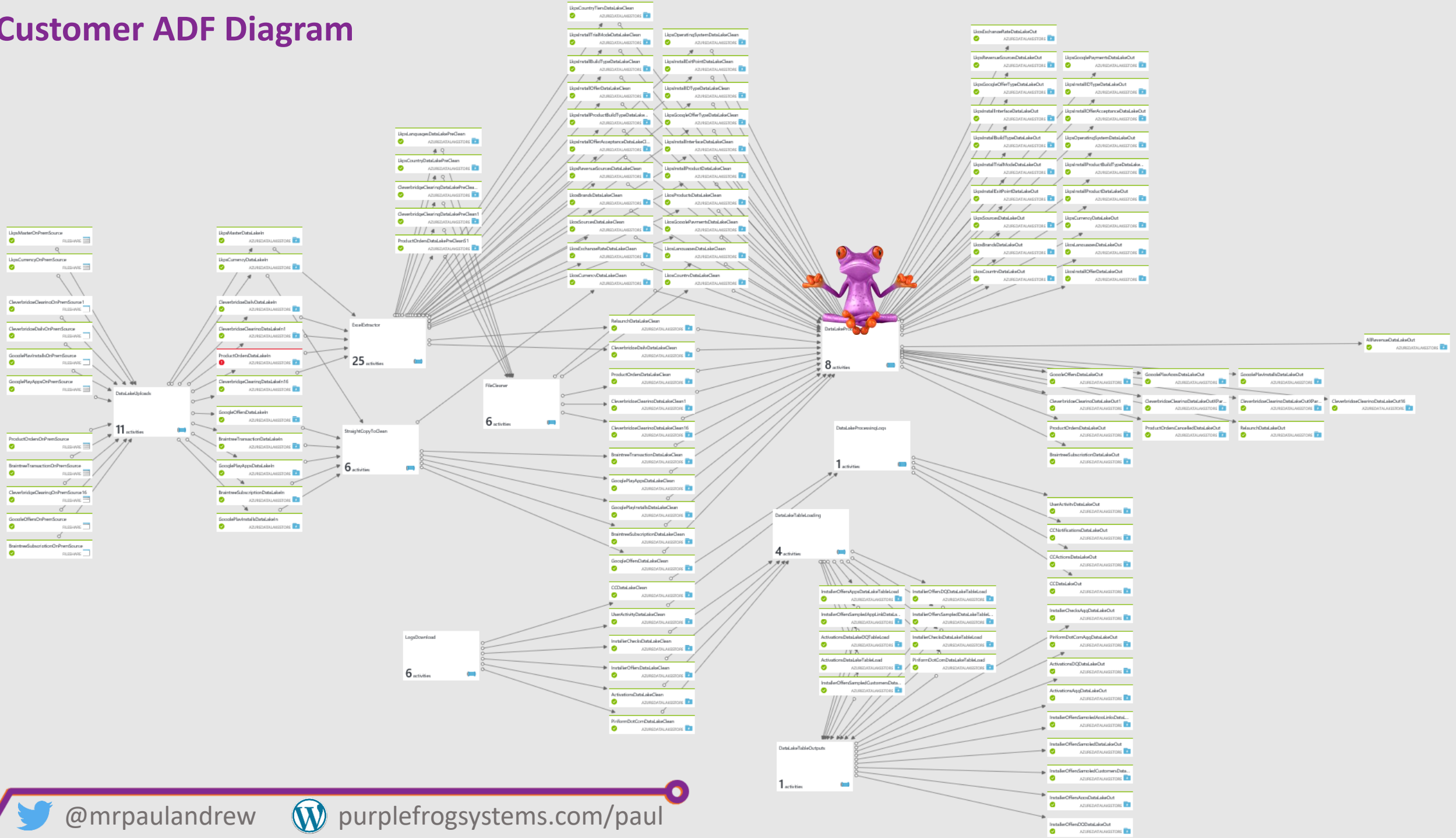
[Resume-AzureRmDataFactoryPipeline]

[Suspend-AzureRmDataFactoryPipeline]

Take Away Points

1. What is Azure Data Factory (ADF)? — Orchestration tool.
   - What is an ADF Pipeline? — Logical container of work.
   - What is an ADF Dataset? — Input/output.
2. What is an ADF Activity? — Instructions for the service(s) invoked.
3. What is an ADF Time Slice? — Way we control executions. Equal across the time line. Bread!
4. What is a ADF Custom Activity? — Extensibility. Call .Net class library. Compute provided by batch service.
5. Don't fear the ADF diagrams!...

Customer ADF Diagram



Thank You