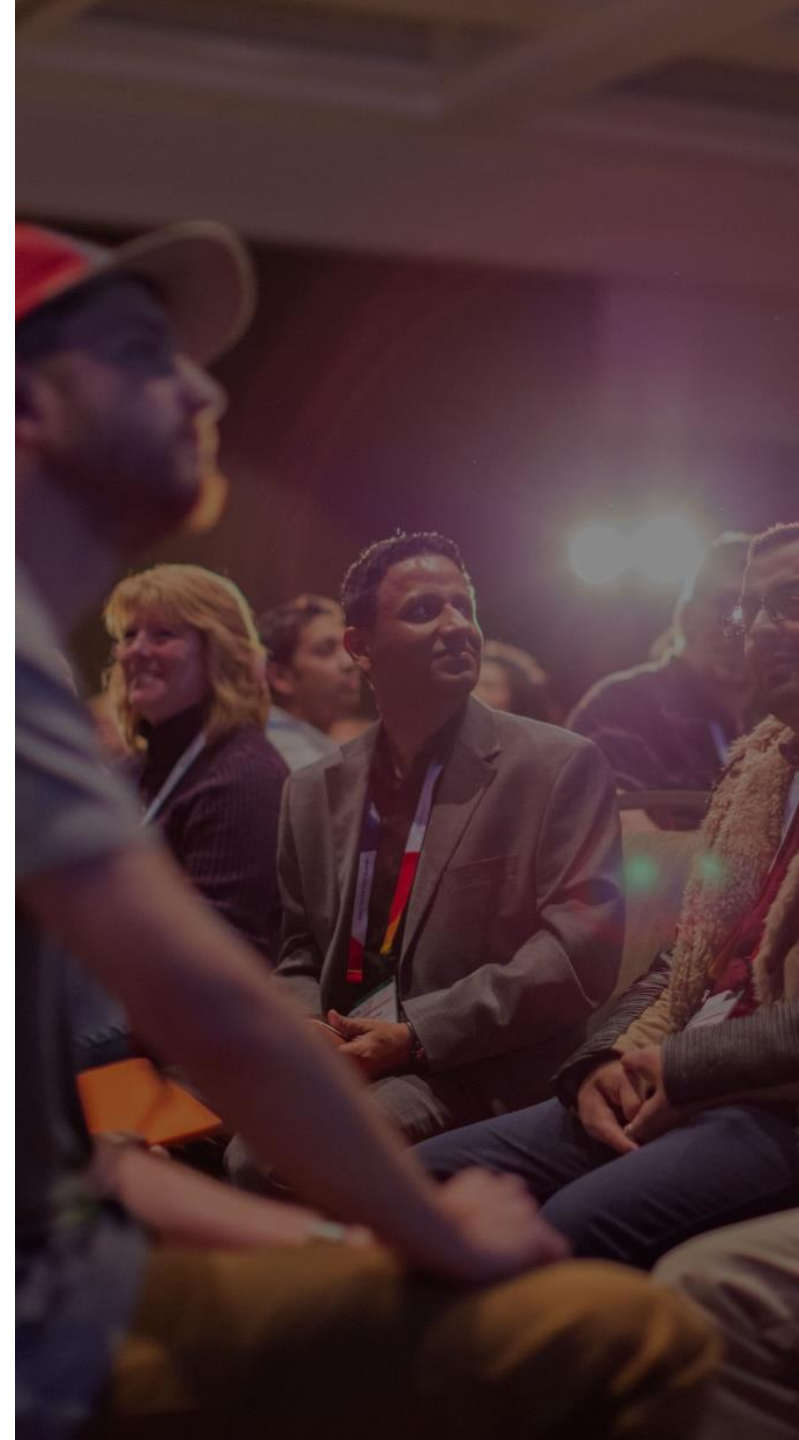




Working With Azure Data Factory & Creating Custom Activities

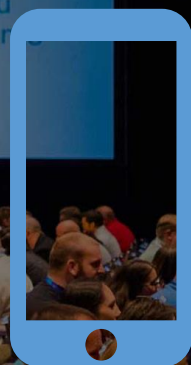
Paul Andrew

Business Intelligence Consultant & Microsoft Data Platform MVP
Purple Frog Systems



QUESTION
#2

What did you
study to go in
your field?



Please silence
cell phones

QUESTION
#2

What did you
study to go into
your field?

Explore everything PASS has to offer



24HOURS
OF
PASS

Free online webinar
events



LOCAL
GROUPS

Local user groups
around the world



SQLSATURDAY
PASS

Free 1-day local training
events



VIRTUAL
GROUPS

Online special interest
user groups



BUSINESS
ANALYTICS DAY
PASS

Business analytics
training



PASS
VOLUNTEERS

Get involved

Free Online Resources

PASS Blog
White Papers
Session Recordings

Newsletter

PASS Connector
BA Insights

www.pass.org

Session evaluations

Your feedback is important and valuable.

Submit by 5pm Friday, November 10th to win prizes. **3 Ways to Access:**



Go to passSummit.com



Download the GuideBook App and
search: PASS Summit 2017



Follow the QR code link displayed
on session signage throughout the
conference venue and in the
program guide



Paul Andrew

Business Intelligence Consultant



• [/mrpaulandrew](https://www.linkedin.com/company/mrpaulandrew)



• [@mrpaulandrew](https://twitter.com/mrpaulandrew)



• purplefrogsystems.com/paul

Many Years' On Premises Experience

10+ years' experience working with the complete on premises SQL Server stack in a variety of roles and industries.

Azure Data Services Consultant

Specialising in Azure Data Lake Analytics, Azure Data Factory, Azure Stream Analytics, Cosmos DB, Power BI, Azure Automation, Event Hubs and IoT.

Stack Overflow Top Answerer

On the Azure Data Factory tag I am the top answerer and have earned the 'Unsung Hero' badge for my community contributions.

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 a custom activity for file cleaning.
10. Q&A



Sorry this is not an ADFv2 session

31st Aug – Given private preview access under MVP NDA.

9th Sept – Presenting ADF mini con at SQL Saturday Cambridge.

13th Sept – PASS Summit presentation upload deadline.

25th Sept – ADFv2 public preview announced at MS Ignite.

27th Sept – ADFv2 blog post done. purplefrogsystems.com/paul

9th to 15th Oct – SQL Relay UK conference.

20th to 28th Oct – Honey moon in Rome.

31st Oct – Arrived in Seattle.



@mrpaulandrew

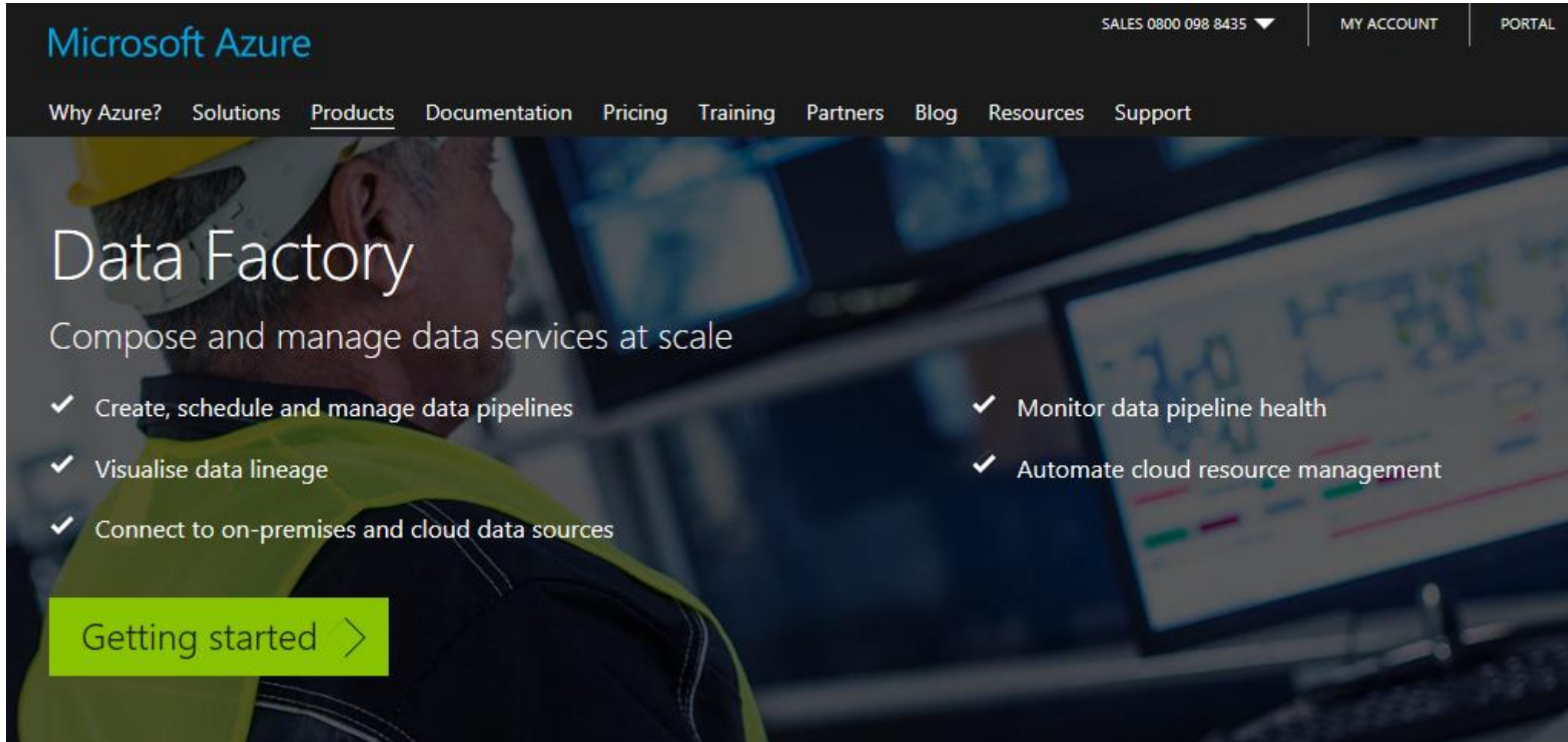


purplefrogsystems.com/paul



What is Azure Data Factory (ADF)?

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



Microsoft Azure

SALES 0800 098 8435 ▼ MY ACCOUNT PORTAL

Why Azure? Solutions Products Documentation Pricing Training Partners Blog Resources Support

Data Factory

Compose and manage data services at scale

- ✓ Create, schedule and manage data pipelines
- ✓ Visualise data lineage
- ✓ Connect to on-premises and cloud data sources
- ✓ Monitor data pipeline health
- ✓ Automate cloud resource management

Getting started >



@mrpaulandrew



purplefrogsystems.com/paul

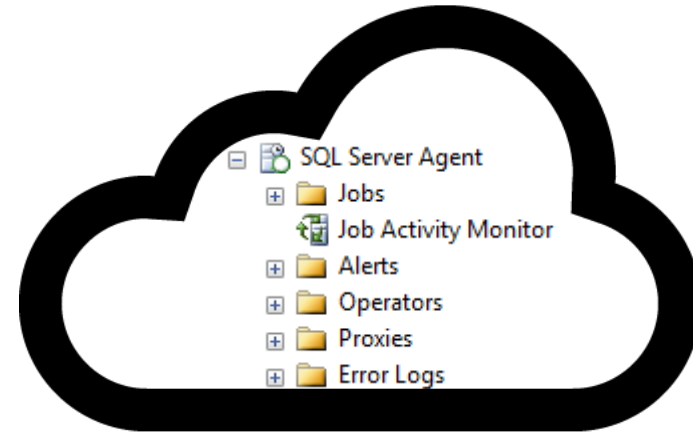


What is Azure Data Factory (ADF)?

What it is...



What it is **NOT**...



@mrpaulandrew



purplefrogsystems.com/paul

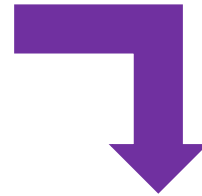
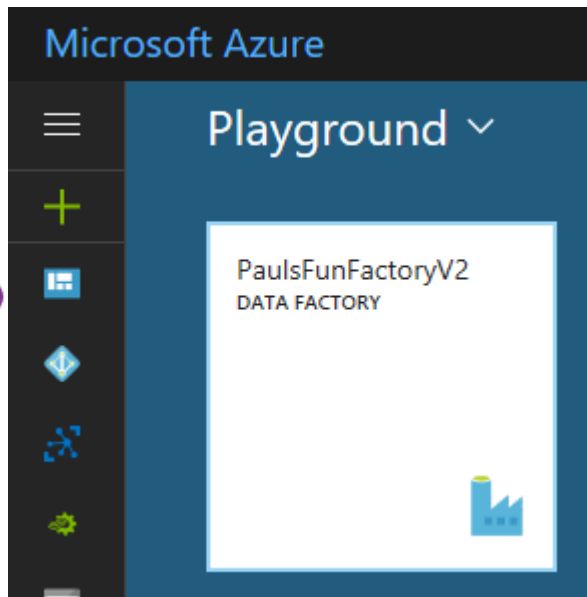


What is an ADF Pipeline and Dataset?



Blog post:

<https://www.purplefrogssystem.com/paul/2017/08/chaining-azure-data-factory-activities-and-datasets/>



1x File



Datasets



Pipelines



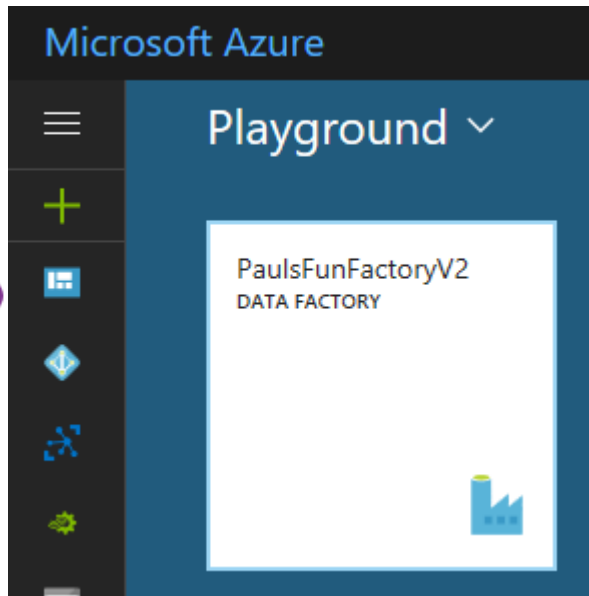
@mrpaulandrew



[purplefrogssystem.com/paul](https://www.purplefrogssystem.com/paul)



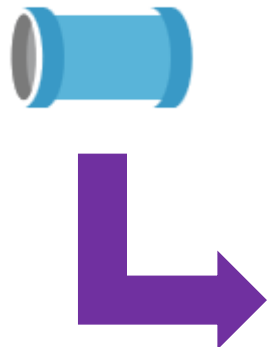
What is an ADF Activity?



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

A screenshot of the Microsoft Azure portal interface. The breadcrumb trail shows 'PaulsFunFactory > PaulsFunFactory > Drafts/Draft-1'. The 'Add activity' button is highlighted with a purple box. Below it, a list of activities is shown:

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



@mrpaulandrew



purplefrogssystems.com/paul

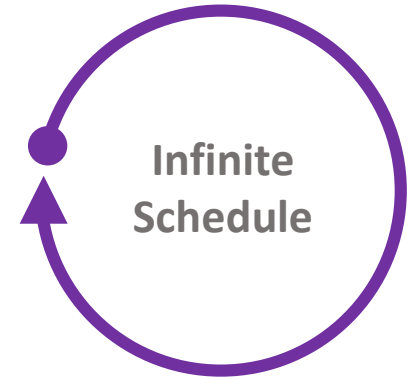


What is an ADF Time Slice?

What it is...



What it is **NOT**...



Triggered Event



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



@mrpaulandrew



purplefrogsystems.com/paul



Time Slice Relationships in ADF and Provisioning

//Input Dataset

```
"availability": {  
  "frequency": "Day",  
  "interval": 1  
}
```

//Activity

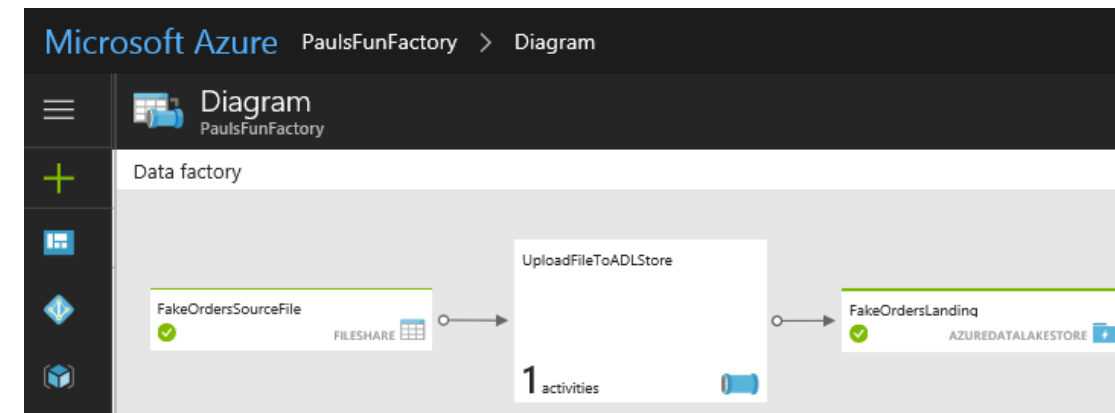
```
"scheduler": {  
  "frequency": "Day",  
  "interval": 1  
}
```

//Pipeline

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

//Output Dataset

```
"availability": {  
  "frequency": "Day",  
  "interval": 1  
}
```



Time Slice Relationships in ADF and Provisioning

//Input Dataset

```
"availability": {  
  "frequency": "Day",  
  "interval": 1  
}
```

//Activity

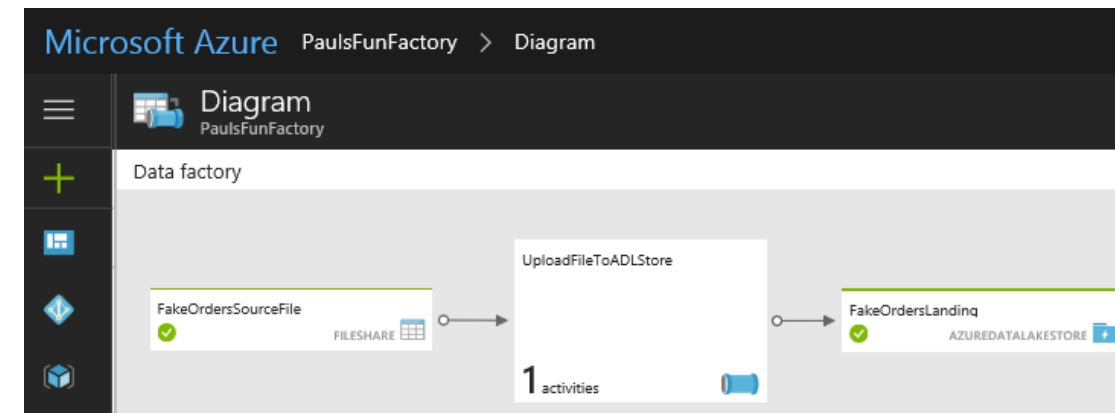
```
"scheduler": {  
  "frequency": "Day",  
  "interval": 1  
}
```

//Pipeline

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

//Output Dataset

```
"availability": {  
  "frequency": "Day",  
  "interval": 1  
}
```



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  
}
```

Data slices (by slice time)
FakeOrdersSourceFile

Filter

SLICE START TIME	SLICE END TIME	STATUS
02/01/2017 12:00 AM UTC (Wed,...	03/01/2017 12:00 AM UTC (3/1/...	Pending validation
01/01/2017 12:00 AM UTC (1/1/2...	02/01/2017 12:00 AM UTC (Wed...	Ready

Data slices (by slice time)
FakeOrdersLanding

Filter

SLICE START TIME	SLICE END TIME	STATUS
02/01/2017 12:00 AM UTC (Wed,...	03/01/2017 12:00 AM UTC (3/1/...	Pending execution
01/01/2017 12:00 AM UTC (1/1/2...	02/01/2017 12:00 AM UTC (Wed...	Pending execution



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?

PaulsFunFactoryV2
DATA FACTORY

Actions

 Author and deploy

 Copy data (PREVIEW)

 Monitor & Manage ...

Copy Data (PaulsFunFactory)

1 Properties
Recurring copy

2 Source
Connection
Dataset

3 Destination















4 Summary

Source data store

Specify the source data store for the copy task. You can use an existing data store connection (or) specify a new data store. Click [HERE](#) to suggest new copy sources or give comments.

FROM EXISTING CONNECTIONS

CONNECT TO A DATA STORE

						
Amazon Redshift	Amazon S3	Azure Blob Storage	Azure Data Lake Store	Azure DocumentDB	Azure SQL Database	Azure SQL Data Warehouse
						
Azure Table Storage	Cassandra	DB2	File System	FTP	HDFS	MongoDB

Playground ▾

PaulsFunFactory V2
DATA FACTORY

Use The Copy Wizard?

Actions



Author and
deploy



Copy data
(PREVIEW)



Monitor &
Manage



...

Copy Data (PaulsFunFactory)

1 Properties

Recurring copy

2 Source



Connection



Dataset

3 Destination

4 Summary

Source data store

Specify the source data store for the copy task. You can use an existing data store connection (or) specify a new data store. Click [HERE](#) to suggest new copy sources or give comments.

NO!

FROM EXISTING CONNECTIONS

CONNECT TO A DATA STORE



Amazon Redshift



Amazon S3



Azure Blob Storage



Azure Data Lake Store



Azure DocumentDB



Azure SQL Database



Azure SQL Data Warehouse



Azure Table Storage



Cassandra



DB2



File System



FTP



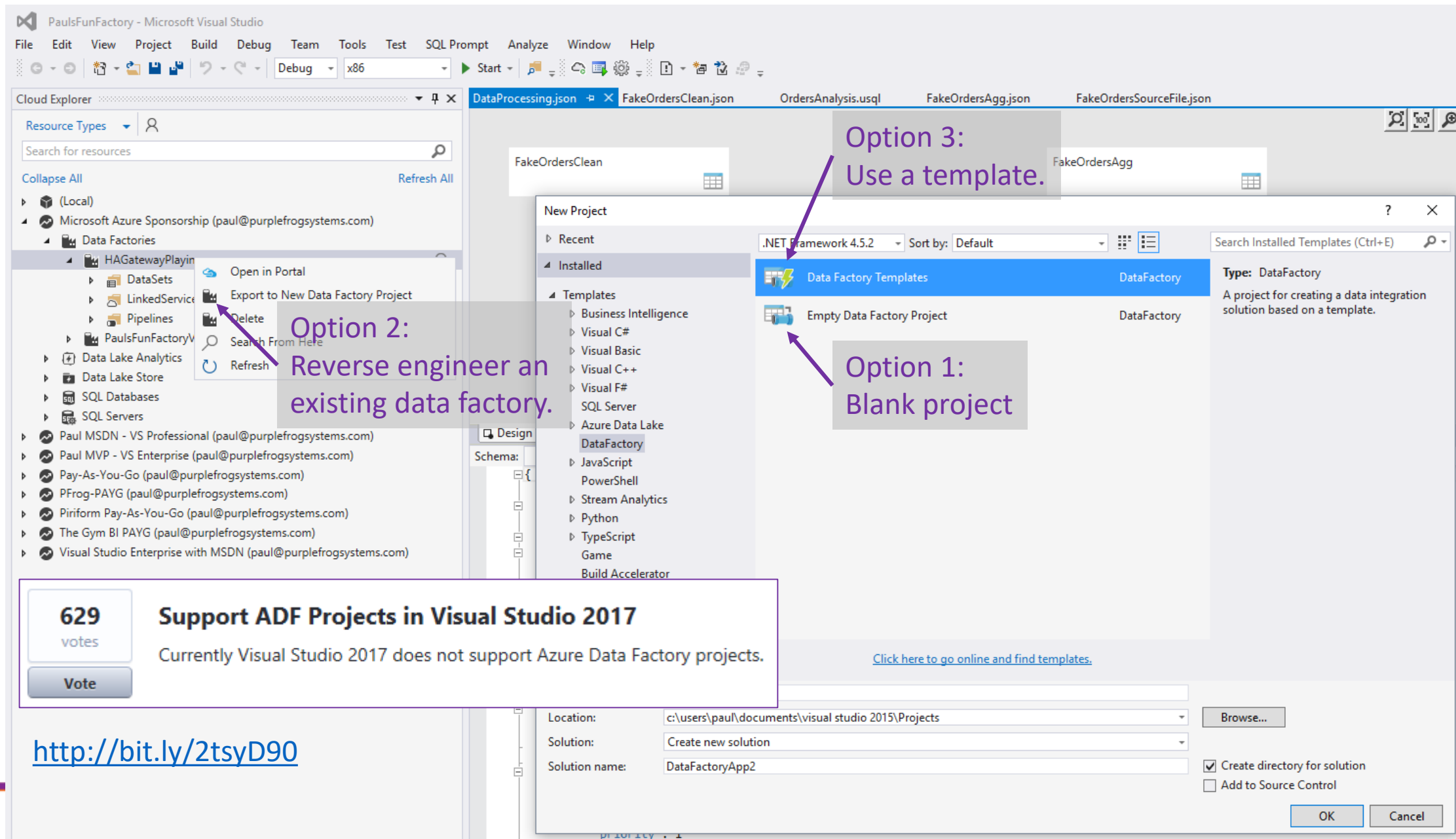
HDFS



MongoDB



Use Visual Studio 2015



Option 3: Use a template.

Option 2: Reverse engineer an existing data factory.

Option 1: Blank project

629 votes

Support ADF Projects in Visual Studio 2017

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

<http://bit.ly/2tsyD90>

Location: c:\users\paul\documents\visual studio 2015\Projects

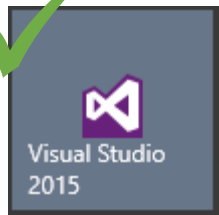
Solution: Create new solution

Solution name: DataFactoryApp2

☒ Create directory for solution

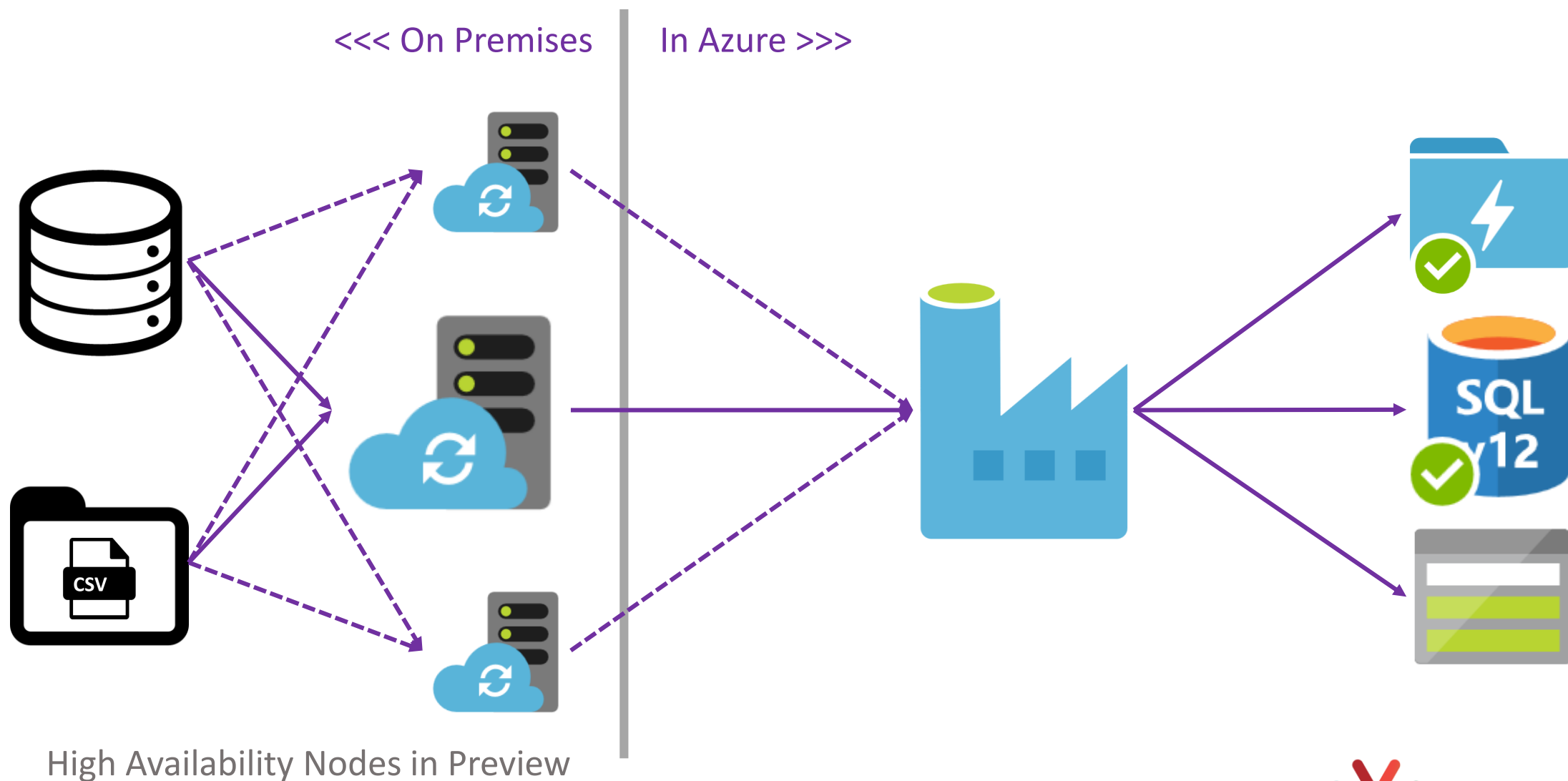
☐ Add to Source Control

OK Cancel



The Data Management Gateway (Integration Runtime)

... nothing to do with Power BI.



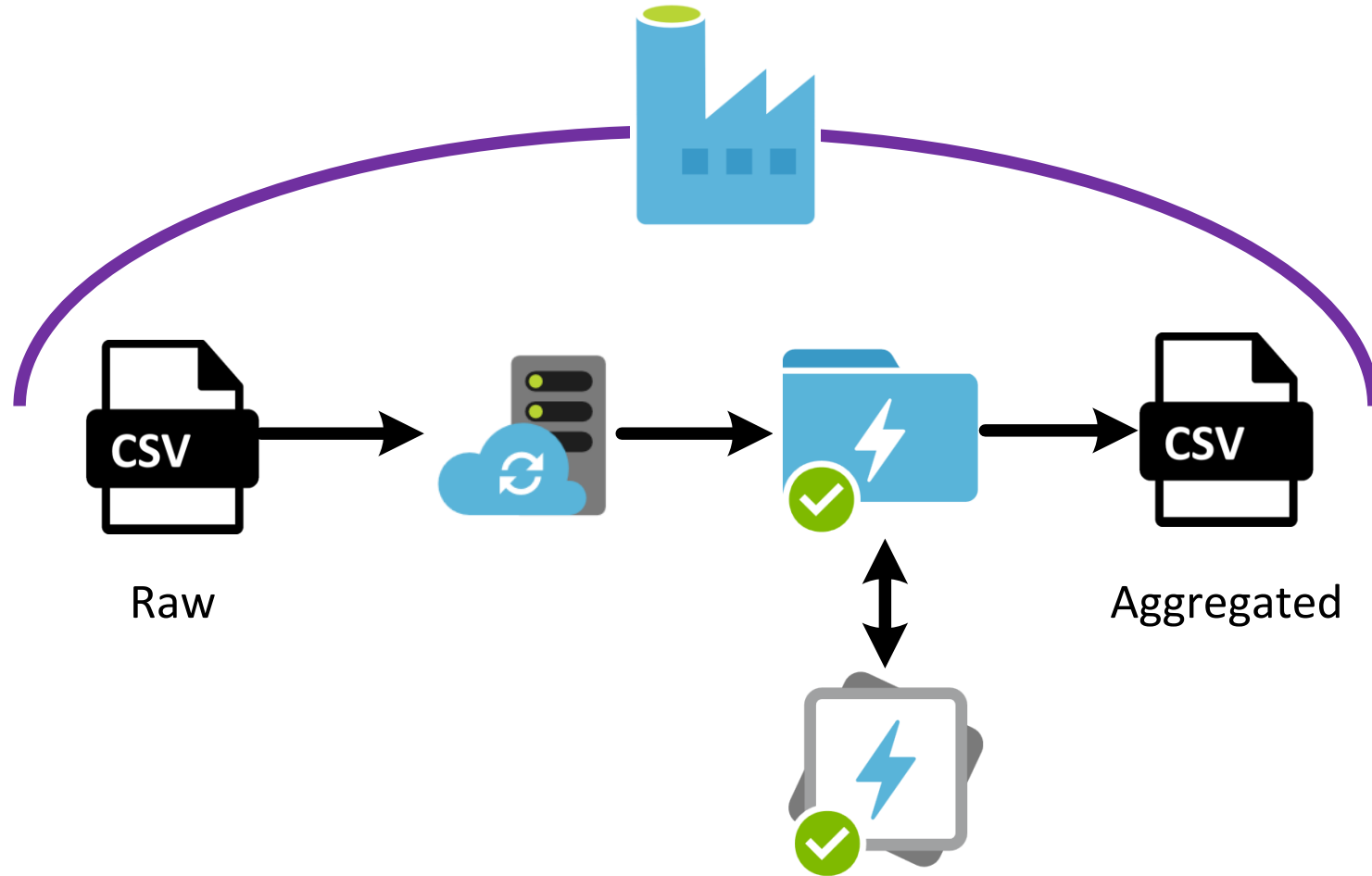
@mrpaulandrew



purplefrogsystems.com/paul



Demo Architecture Version 1



Session Agenda Recap

Part 1

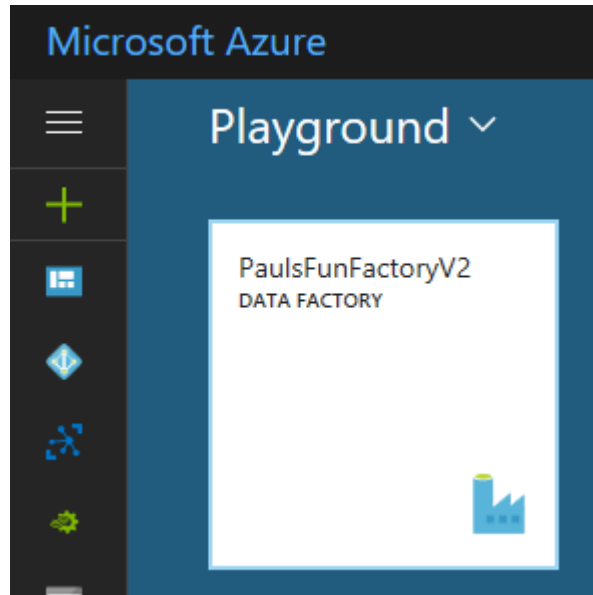
1. What is Azure Data Factory (ADF)?
 - What is an ADF Pipeline?
 - What is an ADF Dataset?
2. What is an ADF Activity?
3. What is an ADF Time Slice?
4. Developer tool for ADF.
5. The ADF Data Management Gateway.
6. Building a basic copy data factory.

Part 2

7. What is a ADF Custom Activity?
8. Creating a custom activity.
9. Using a custom activity for file cleaning.
10. Q&A

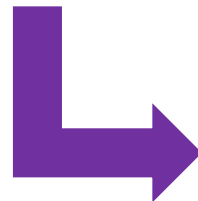


What is an ADF Custom Activity?



Blog post:

<https://www.purplefrogssystem.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": [  
          {
```



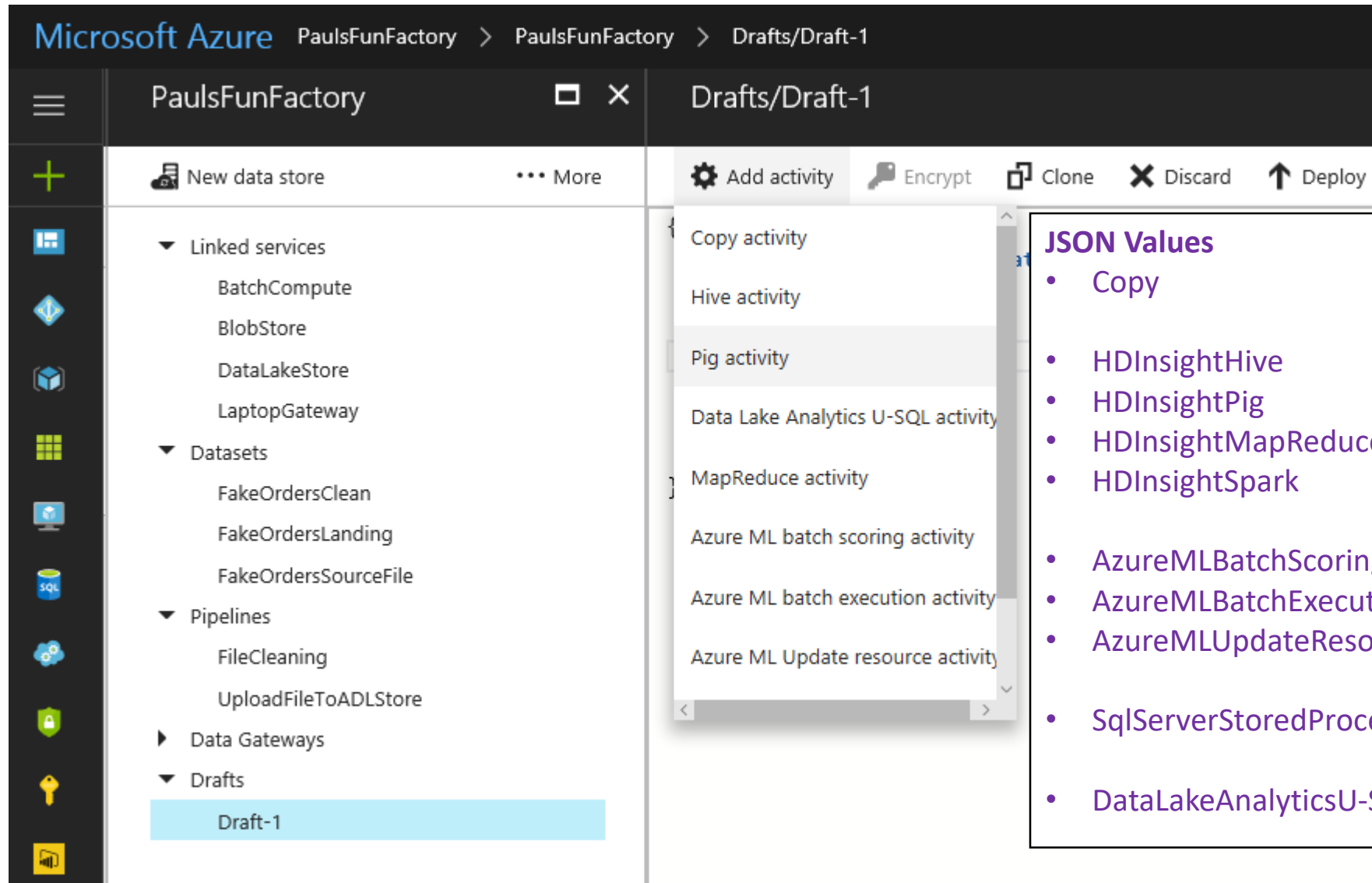
@mrpaulandrew



purplefrogssystem.com/paul



How to Create an ADF Custom Activity?



Microsoft Azure PaulsFunFactory > PaulsFunFactory > Drafts/Draft-1

PaulsFunFactory Drafts/Draft-1

New data store ... More

Add activity Encrypt Clone Discard Deploy

Copy activity

Hive activity

Pig activity

Data Lake Analytics U-SQL activity

MapReduce activity

Azure ML batch scoring activity

Azure ML batch execution activity

Azure ML Update resource activity

Linked services

- BatchCompute
- BlobStore
- DataLakeStore
- LaptopGateway

Datasets

- FakeOrdersClean
- FakeOrdersLanding
- FakeOrdersSourceFile

Pipelines

- FileCleaning
- UploadFileToADLStore

Data Gateways

Drafts

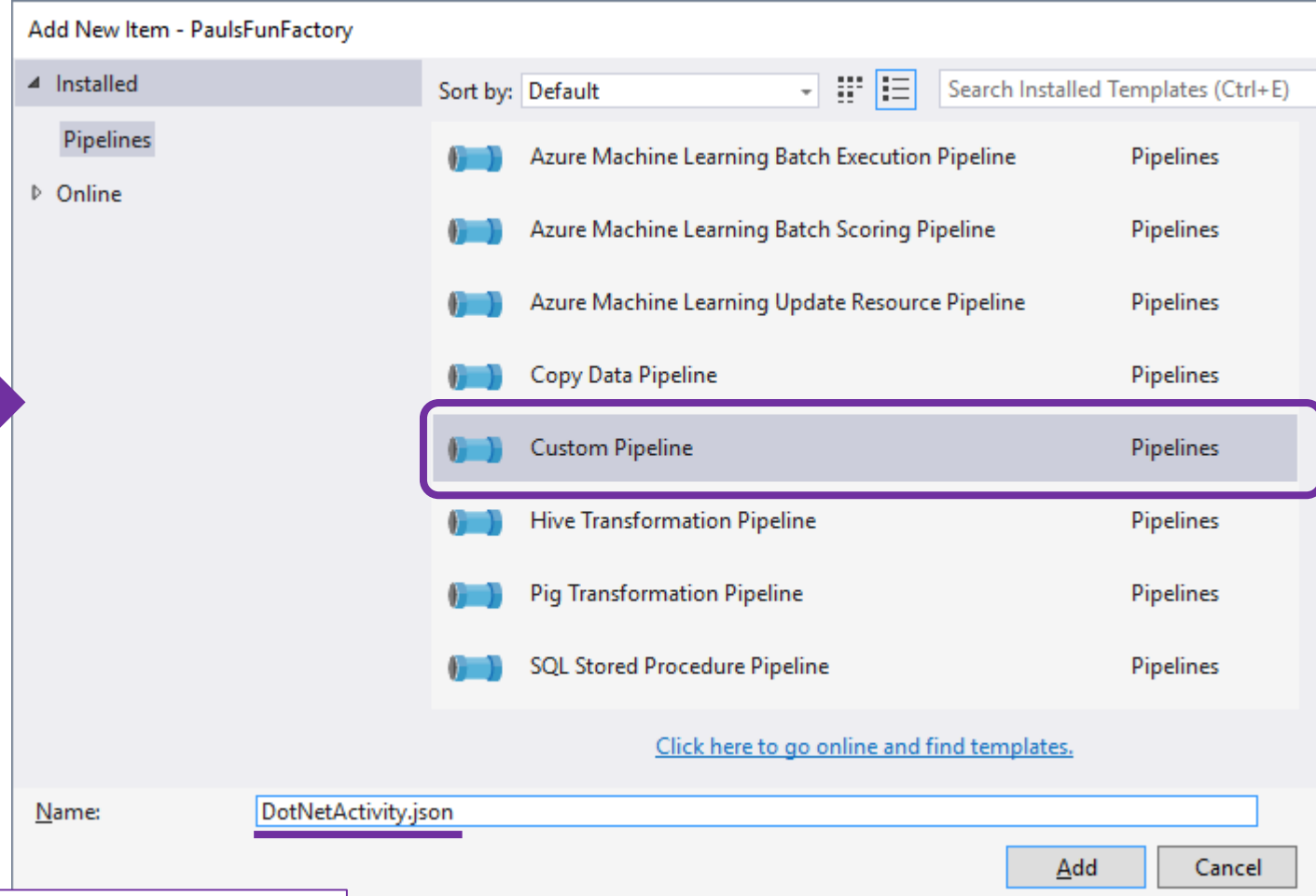
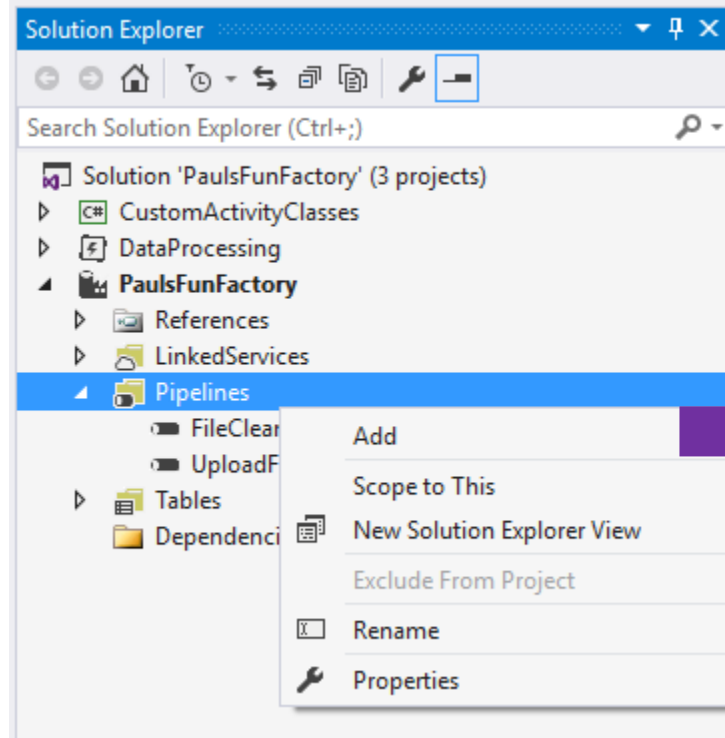
- Draft-1

JSON Values

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



How to Create an ADF Custom Activity?



<http://bit.ly/2tsyD90>

629

votes

Vote

Support ADF Projects in Visual Studio 2017

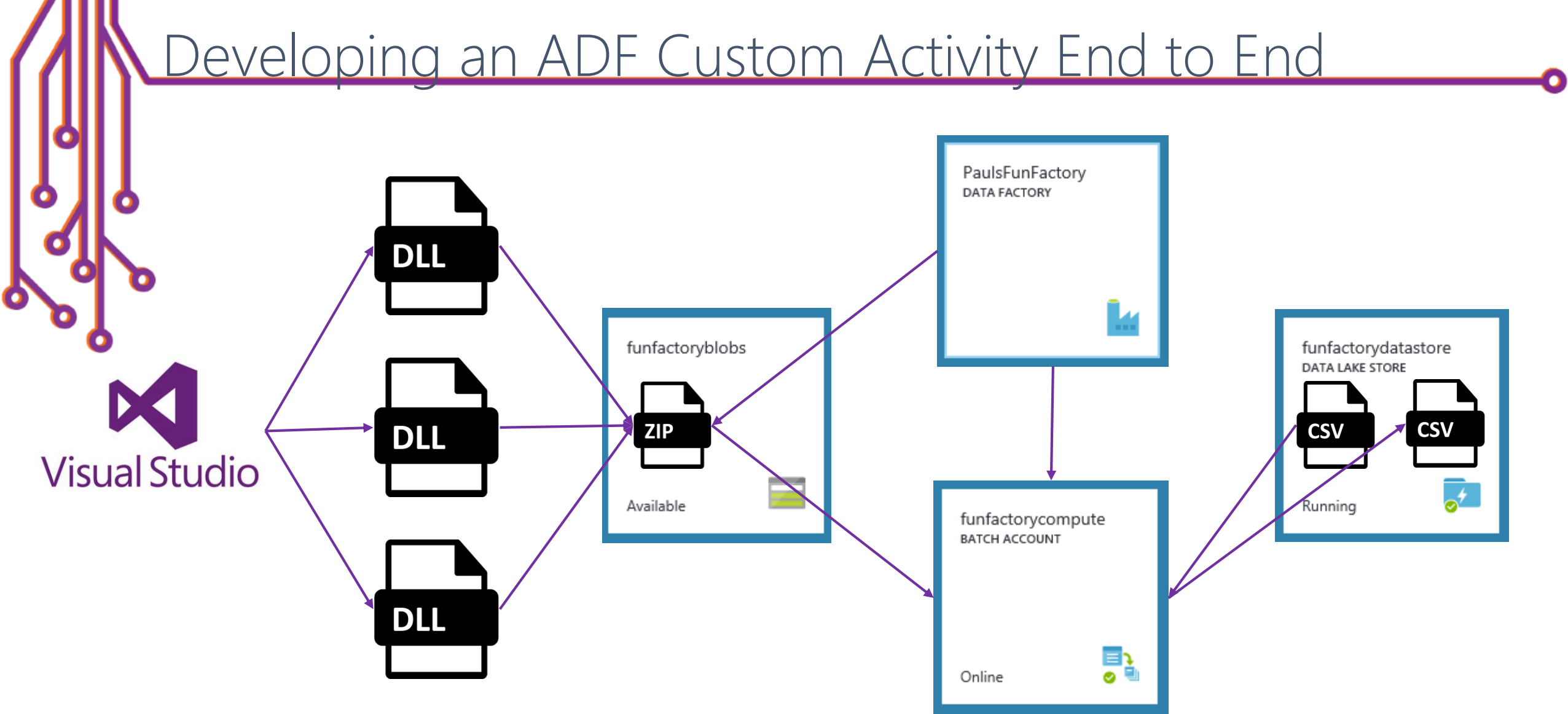
Currently Visual Studio 2017 does not support Azure Data Factory projects.

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

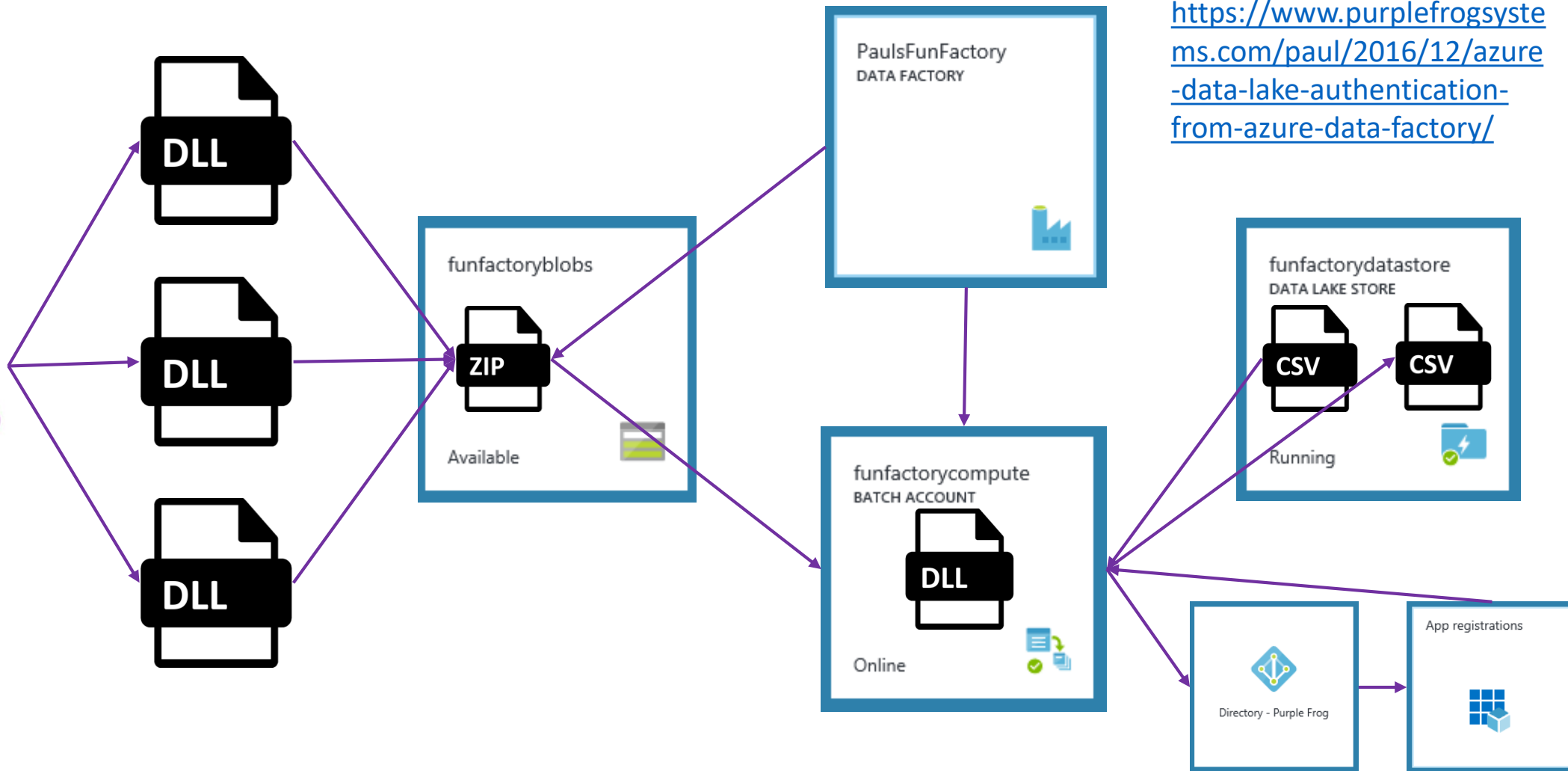


Developing an ADF Custom Activity End to End



Blog post:

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



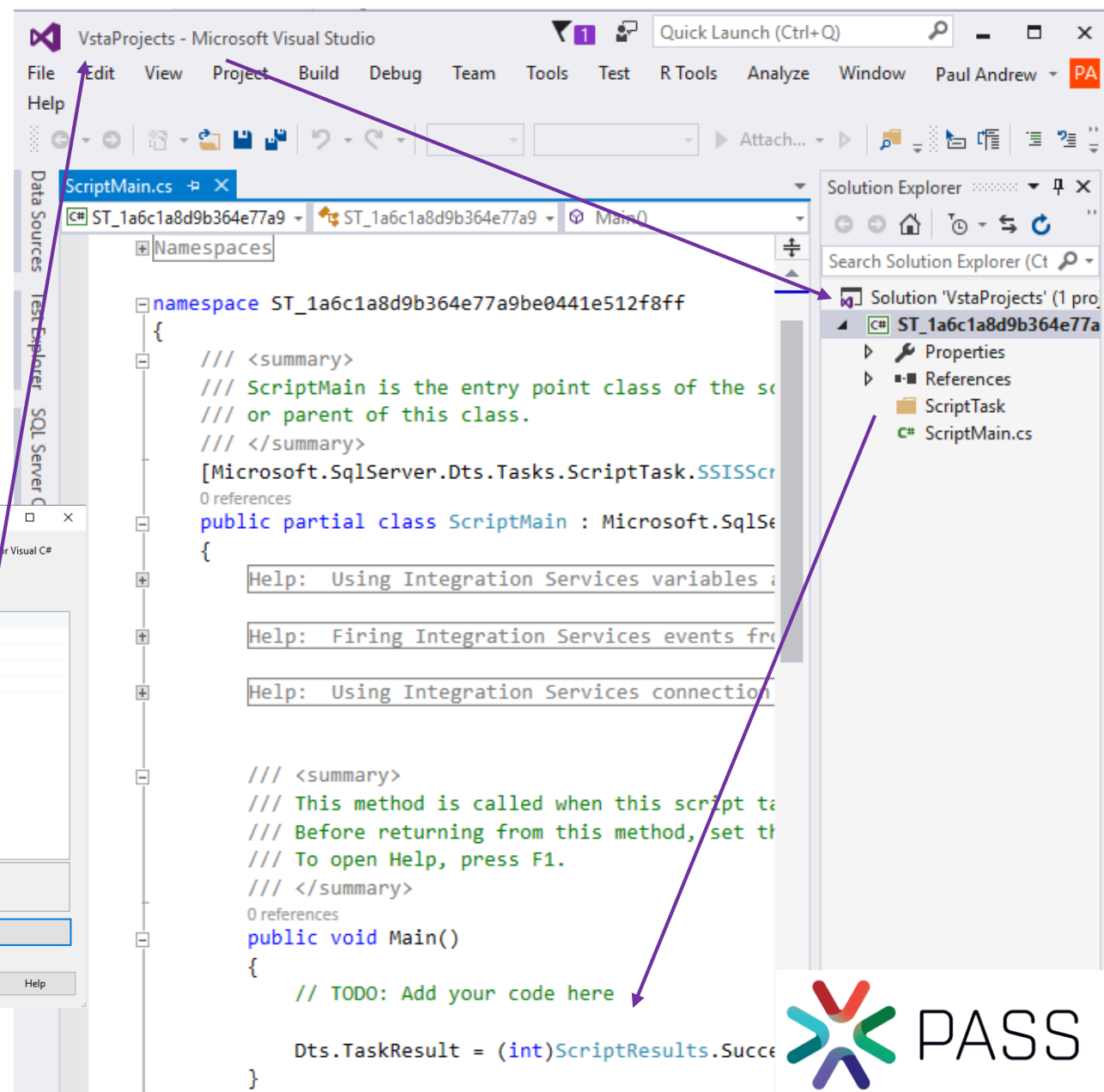
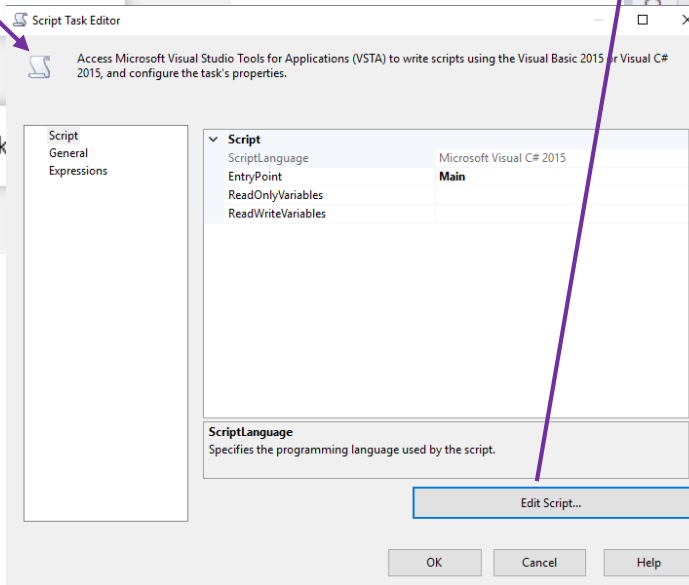
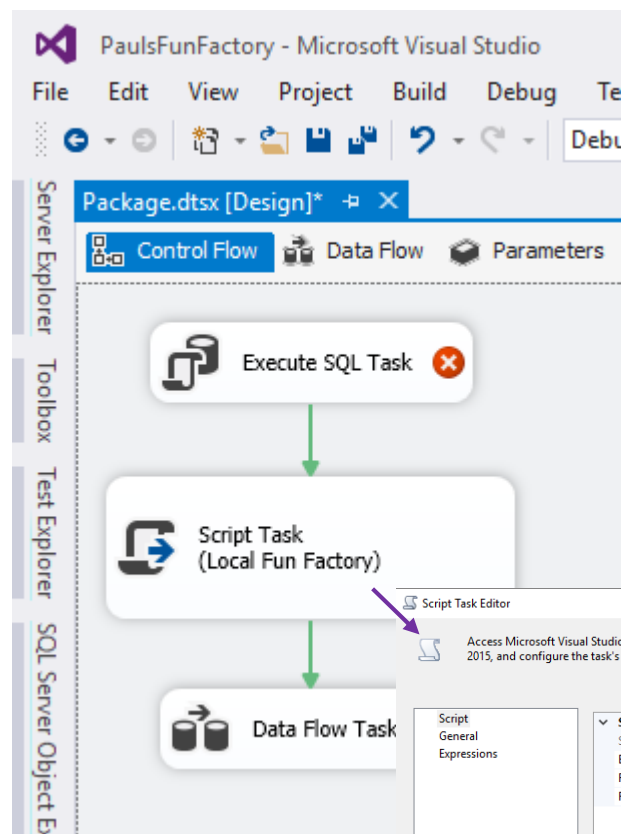
@mrpaulandrew



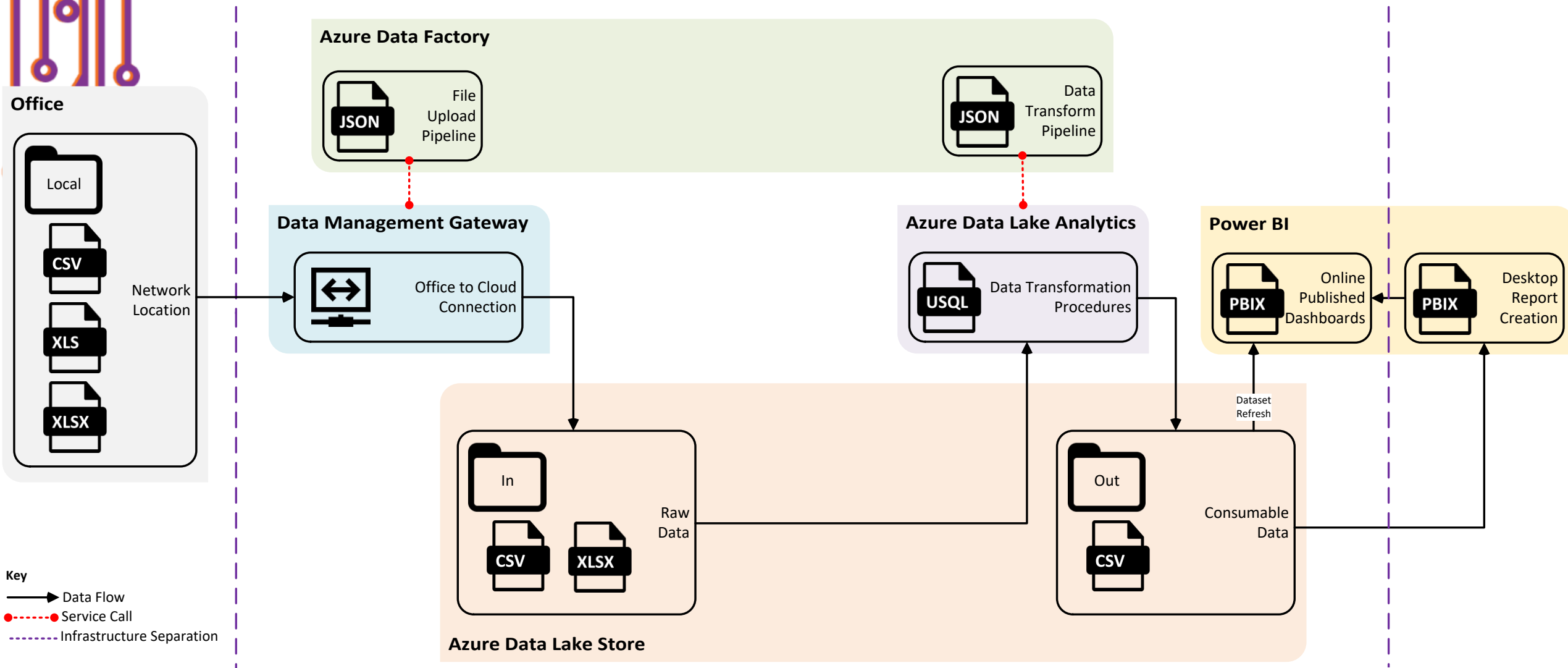
purplefrogsystems.com/paul



ADF Custom Activity vs SSIS Control Flow Script Task



Why do we need an ADF Custom Activity?



* ELT – Extract Load Transform



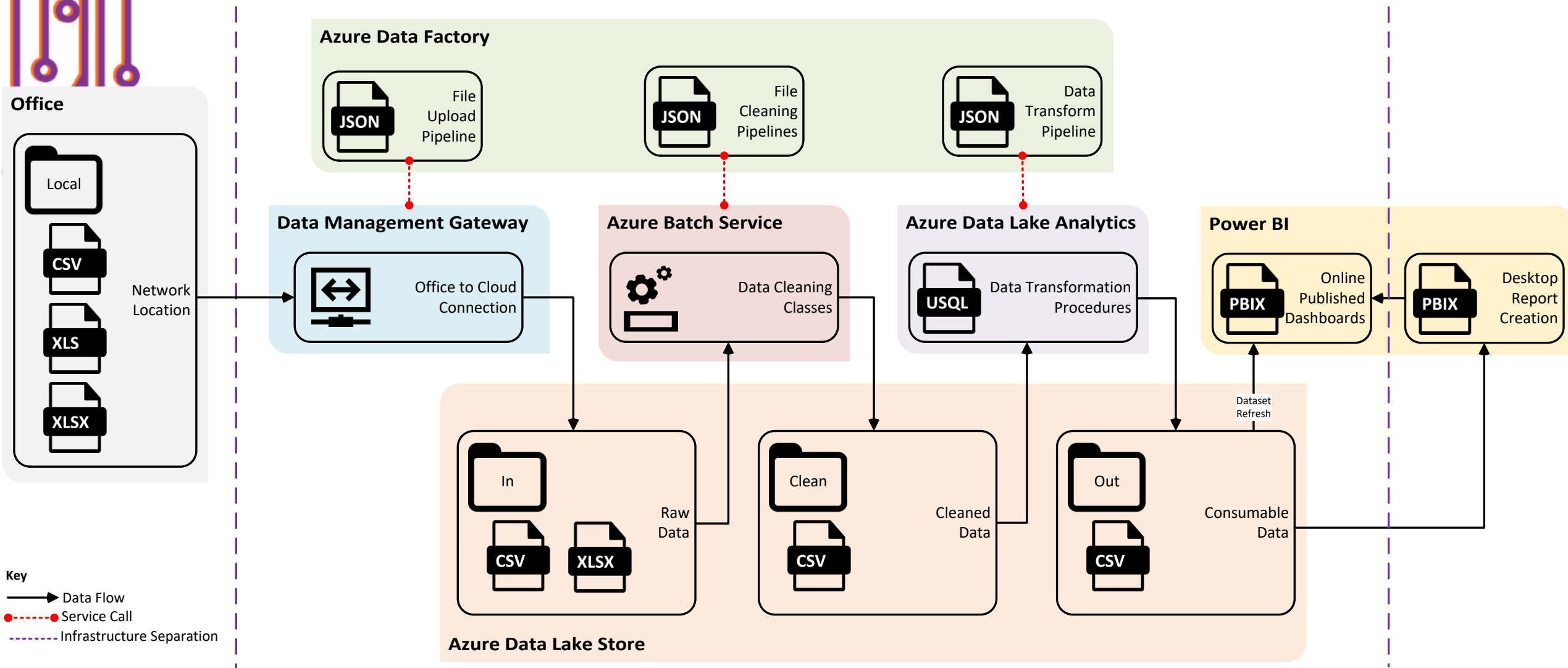
@mrpaulandrew



purplefrogsystems.com/paul

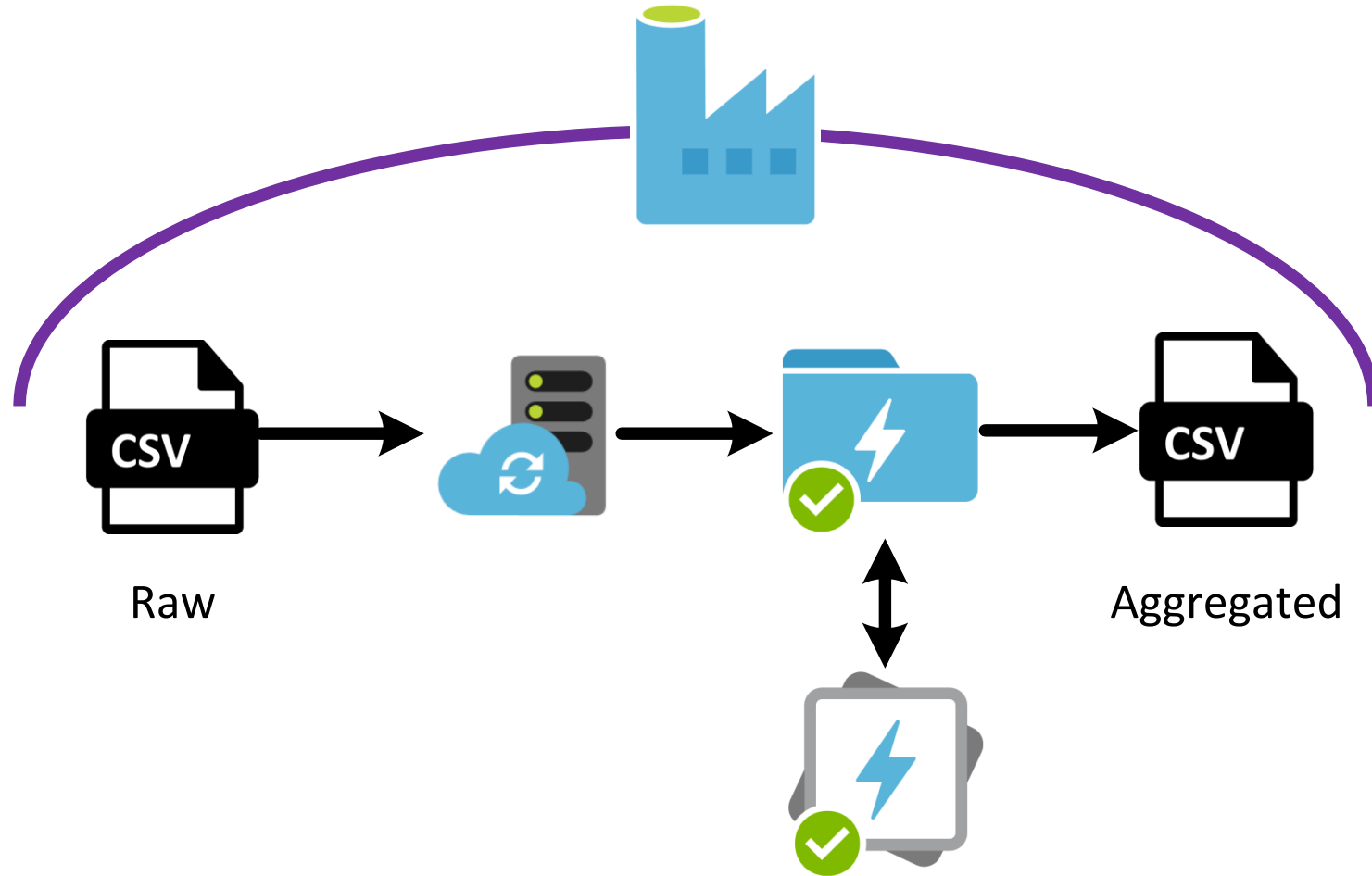


Why do we need an ADF Custom Activity?

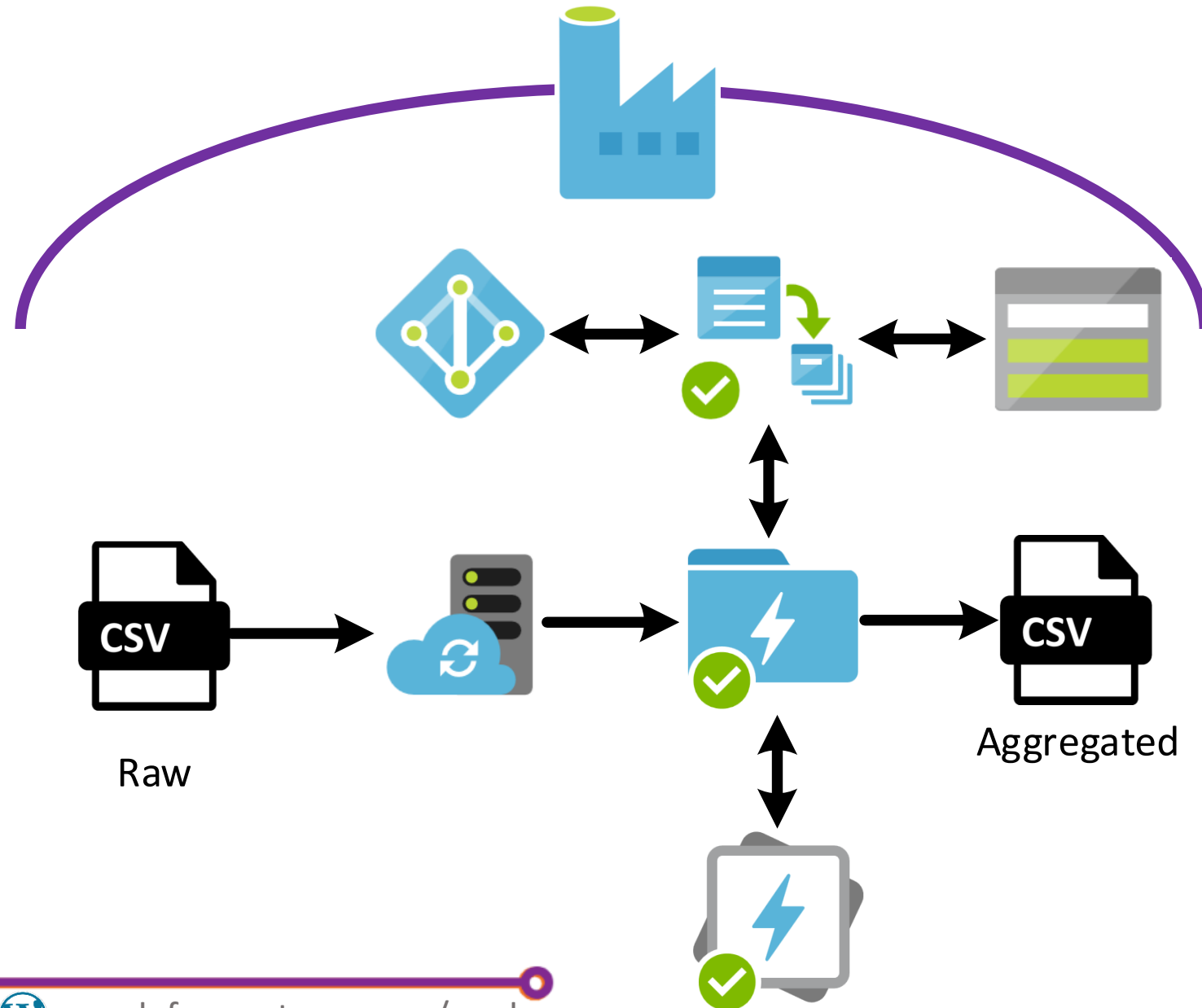


* ECLT – Extract Clean Load Transform

Demo Architecture Version 1



Demo Architecture Version 2



Take Away Points

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. What is a ADF Custom Activity?
5. Don't fear the ADF diagrams!...

Orchestration tool.

Logical container of work.

Input/output.

Instructions for the service(s) invoked.

Way we control executions.

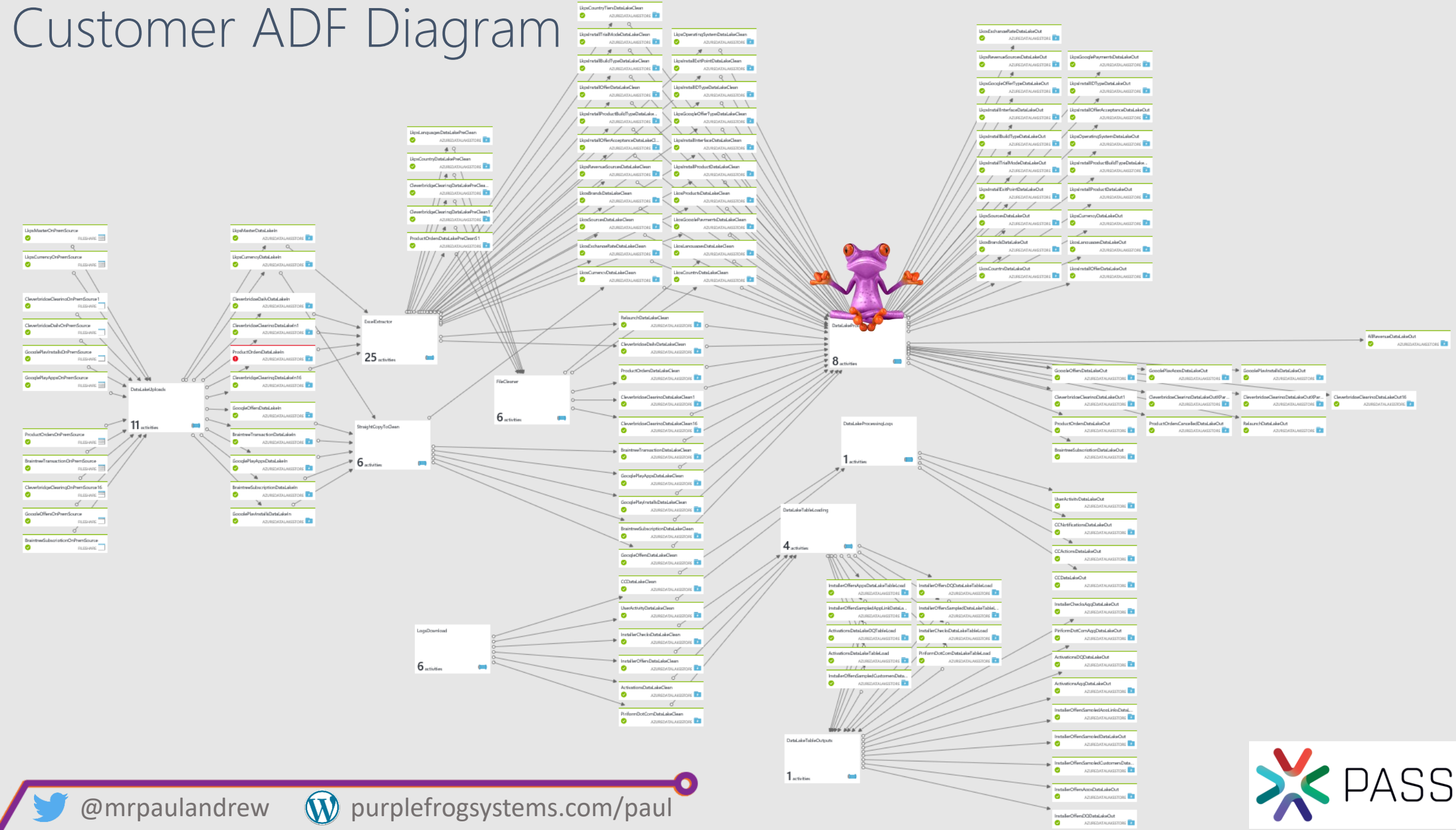
Equal across the time line. Bread!

Extensibility. Call .Net class library.

Compute provided by batch service.



Customer ADF Diagram





Thank You



• [@mrpaulandrew](https://twitter.com/mrpaulandrew)



• paul@purplefrogsystems.com

