Azure Data Lake. The Services. The U-SQL. The C-Sharp

With Paul Andrew

at SQL Ireland UG

Paul Andrew

About Me

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.

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

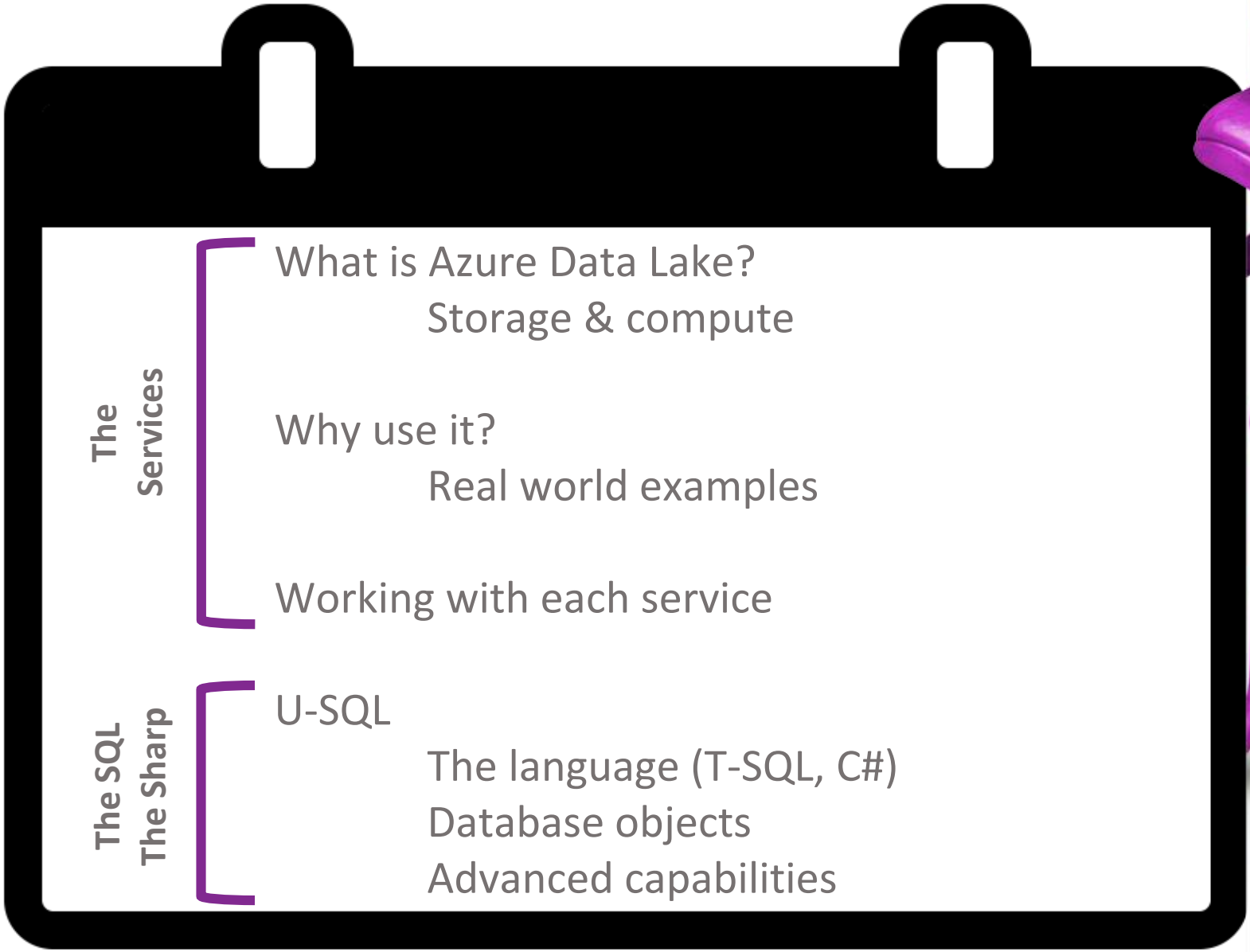
Speaker & helper at various SQL Saturday 's

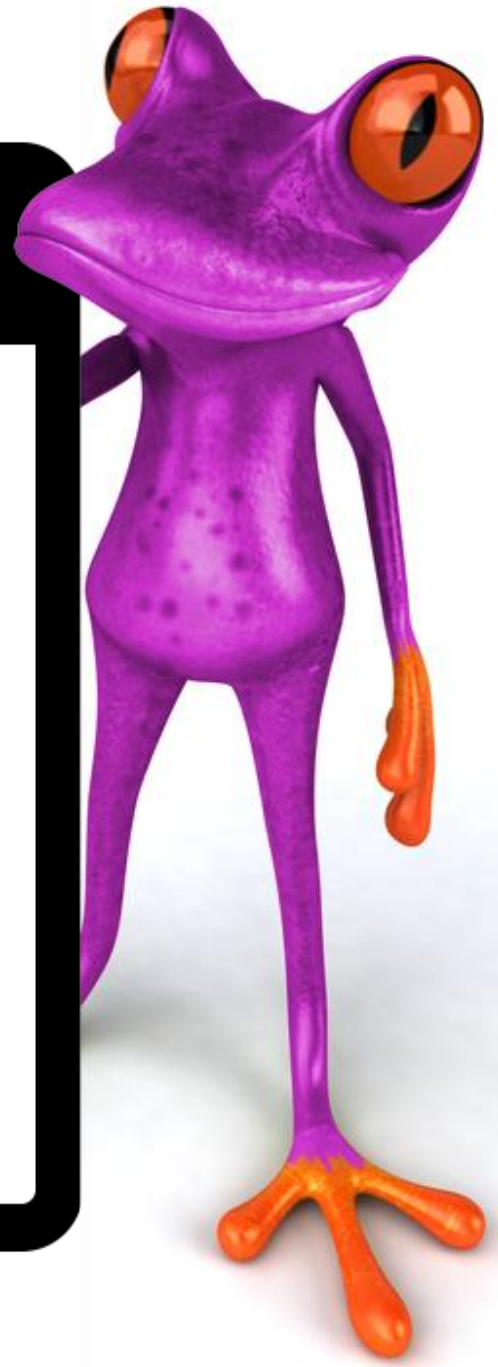
Stack Overflow

Best New Speaker SQL Saturday Dublin 2017

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

Session Agenda



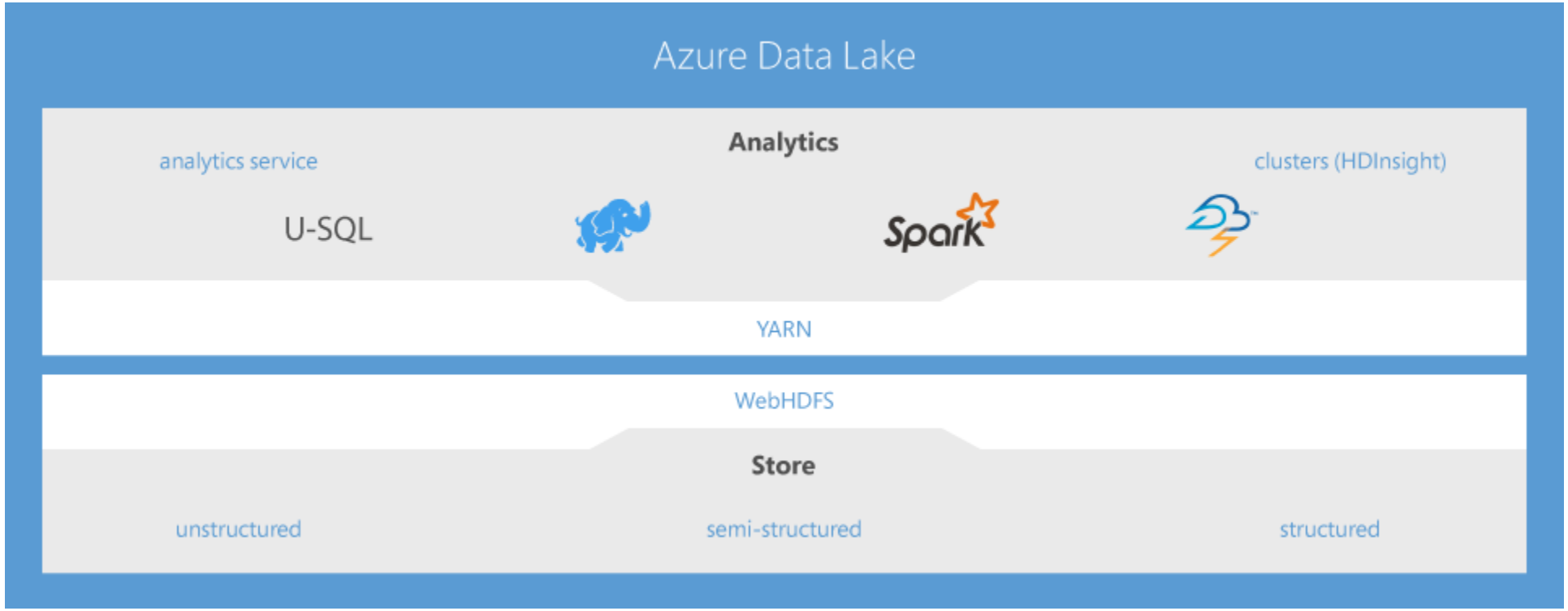


Session Agenda

- **What is Azure Data Lake?**
  - Storage & compute
- Why use it?
  - Real world examples
- Working with each service
- U-SQL
  - The language (T-SQL, C#)
  - Database objects
  - Advanced capabilities

What is Azure Data Lake?

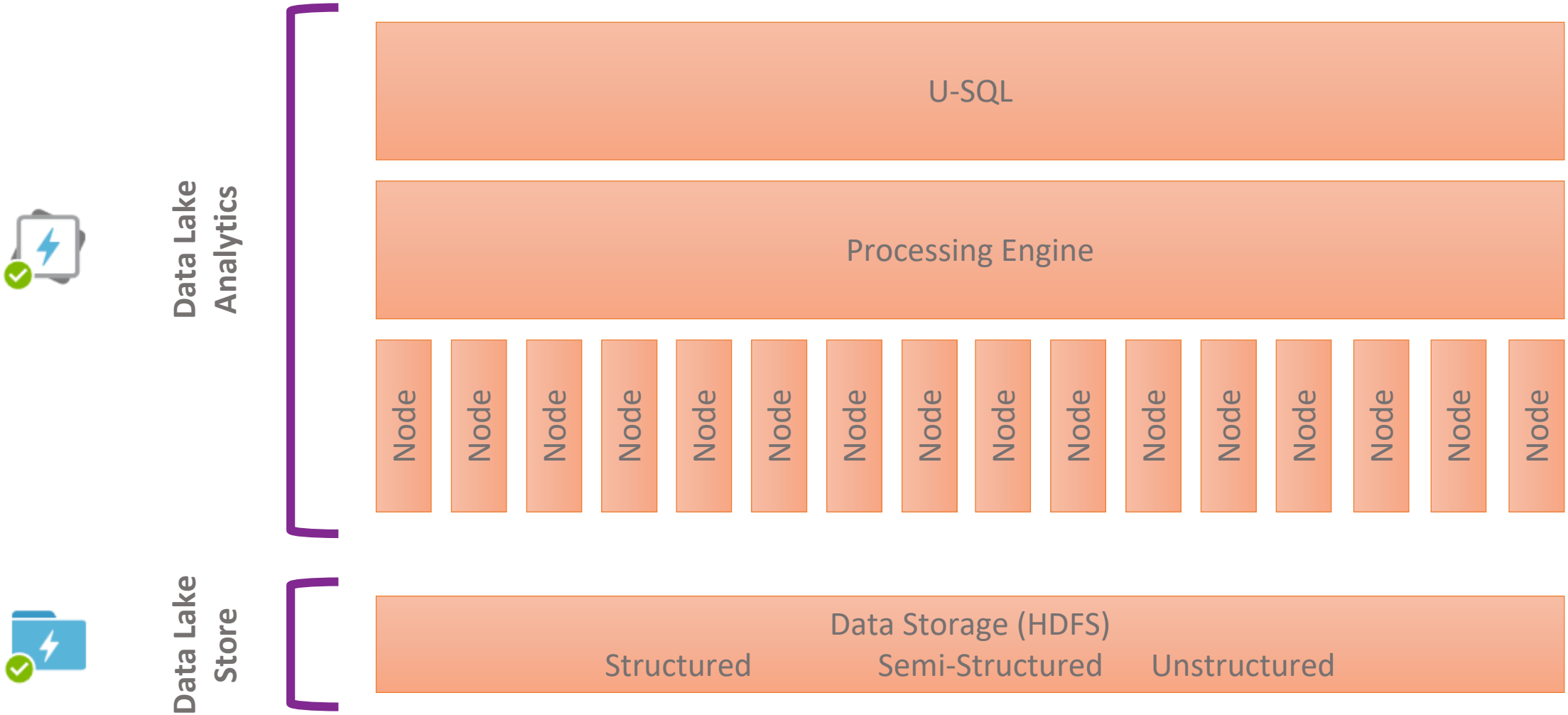
The Microsoft version…



https://azure.microsoft.com/en-gb/solutions/data-lake/

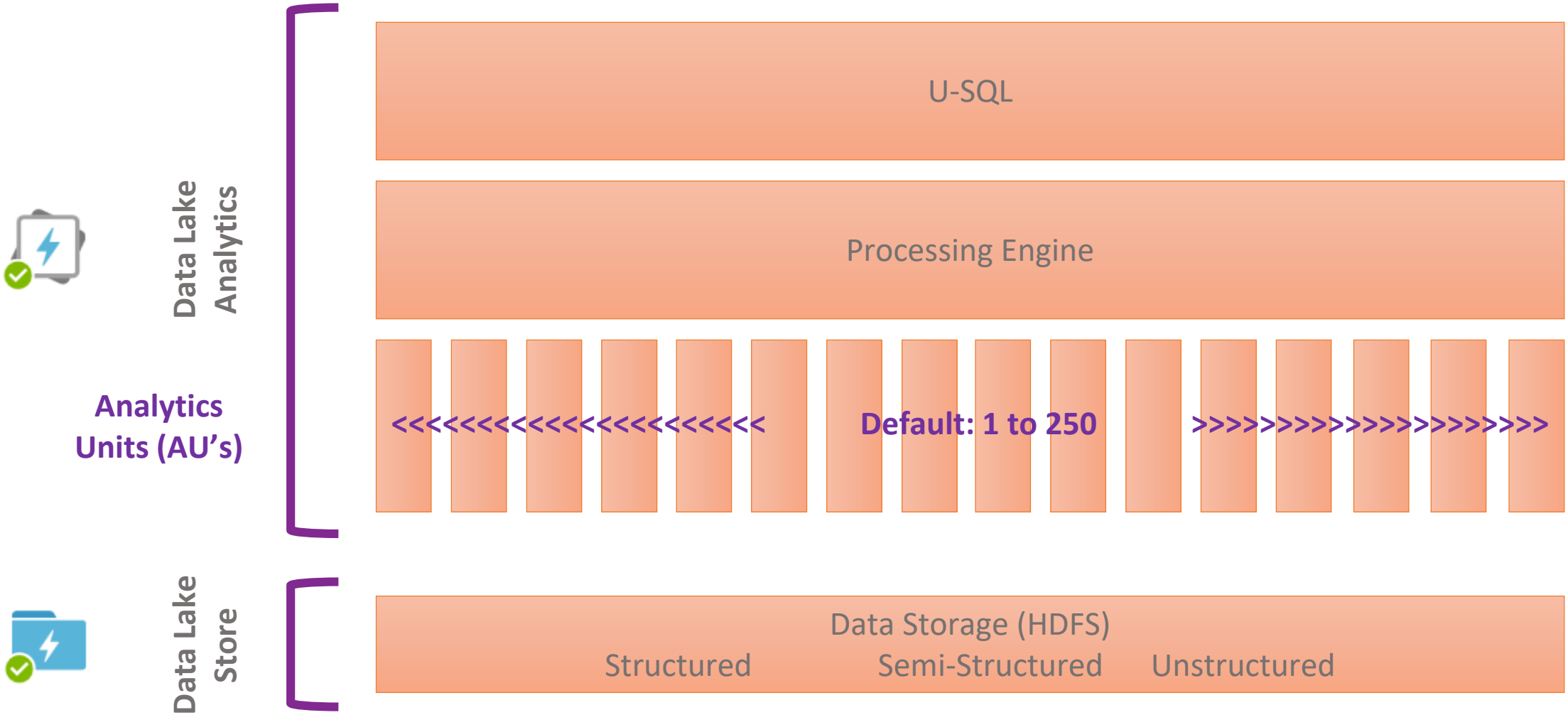
What is Azure Data Lake?

The non product orientated version…



What is Azure Data Lake?

The non product orientated version…



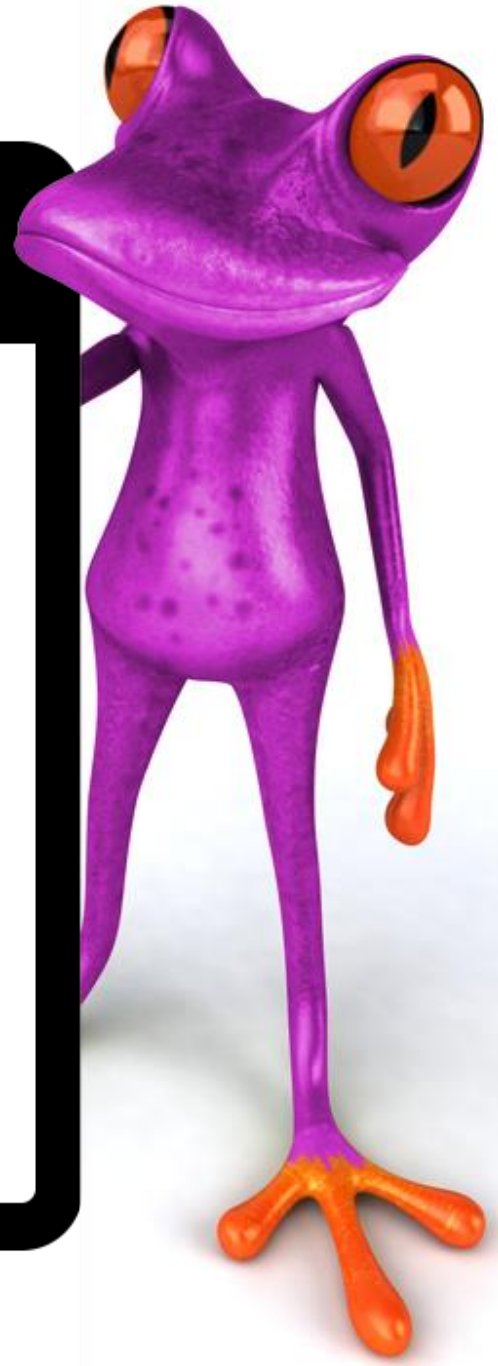
Session Agenda

What is Azure Data Lake?
- Storage & compute

Why use it?
- Real world examples

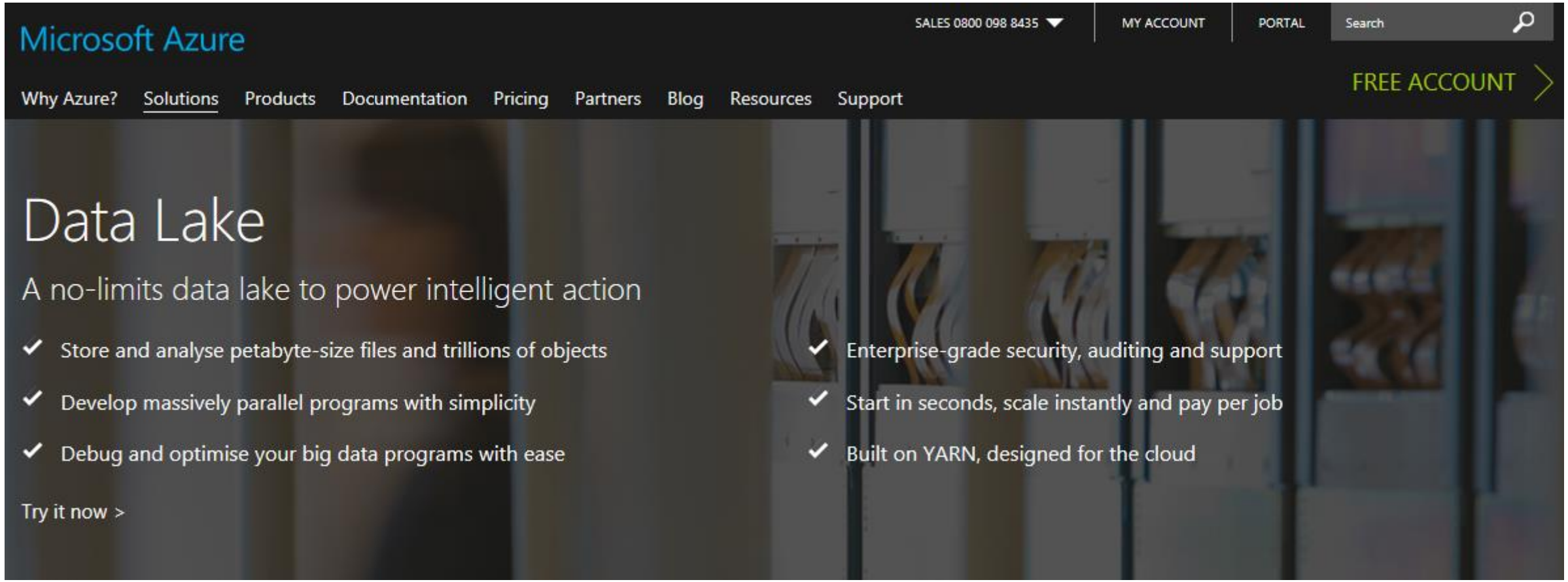
Working with each service

U-SQL
- The language (T-SQL, C#)
- Database objects
- Advanced capabilities



Why Use Azure Data Lake?

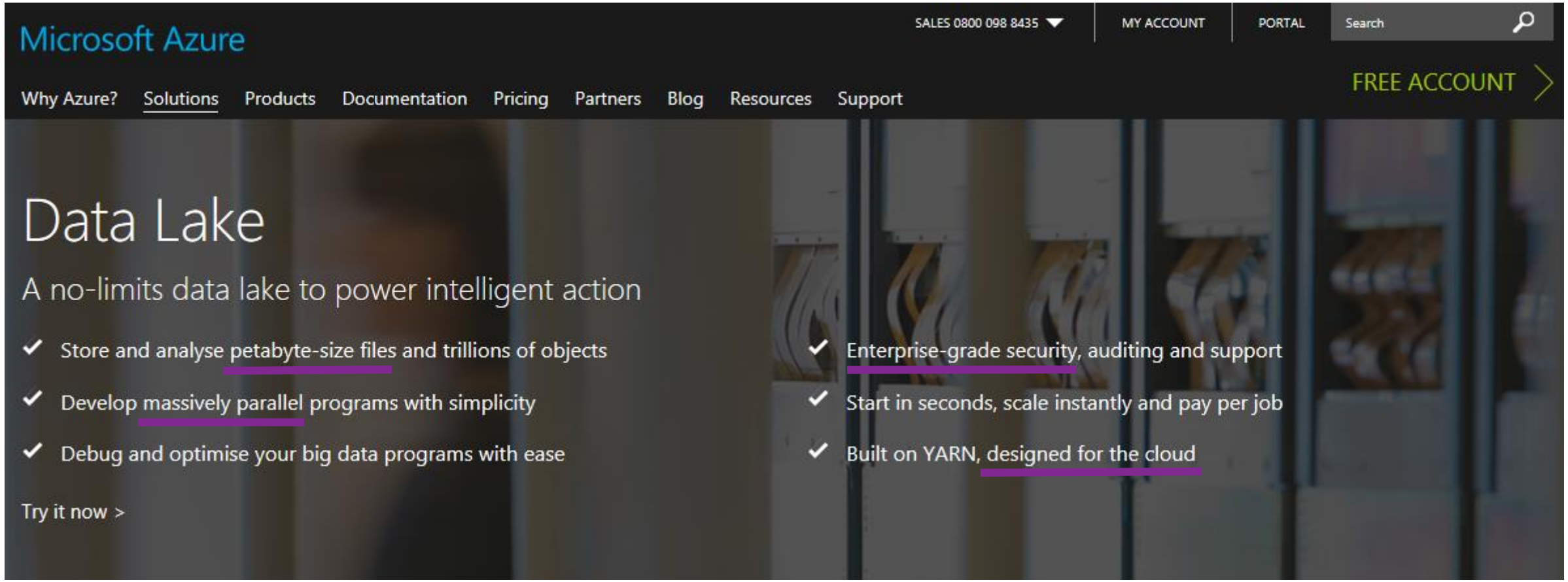
The Microsoft version…



https://azure.microsoft.com/en-gb/solutions/data-lake/

Why Use Azure Data Lake?

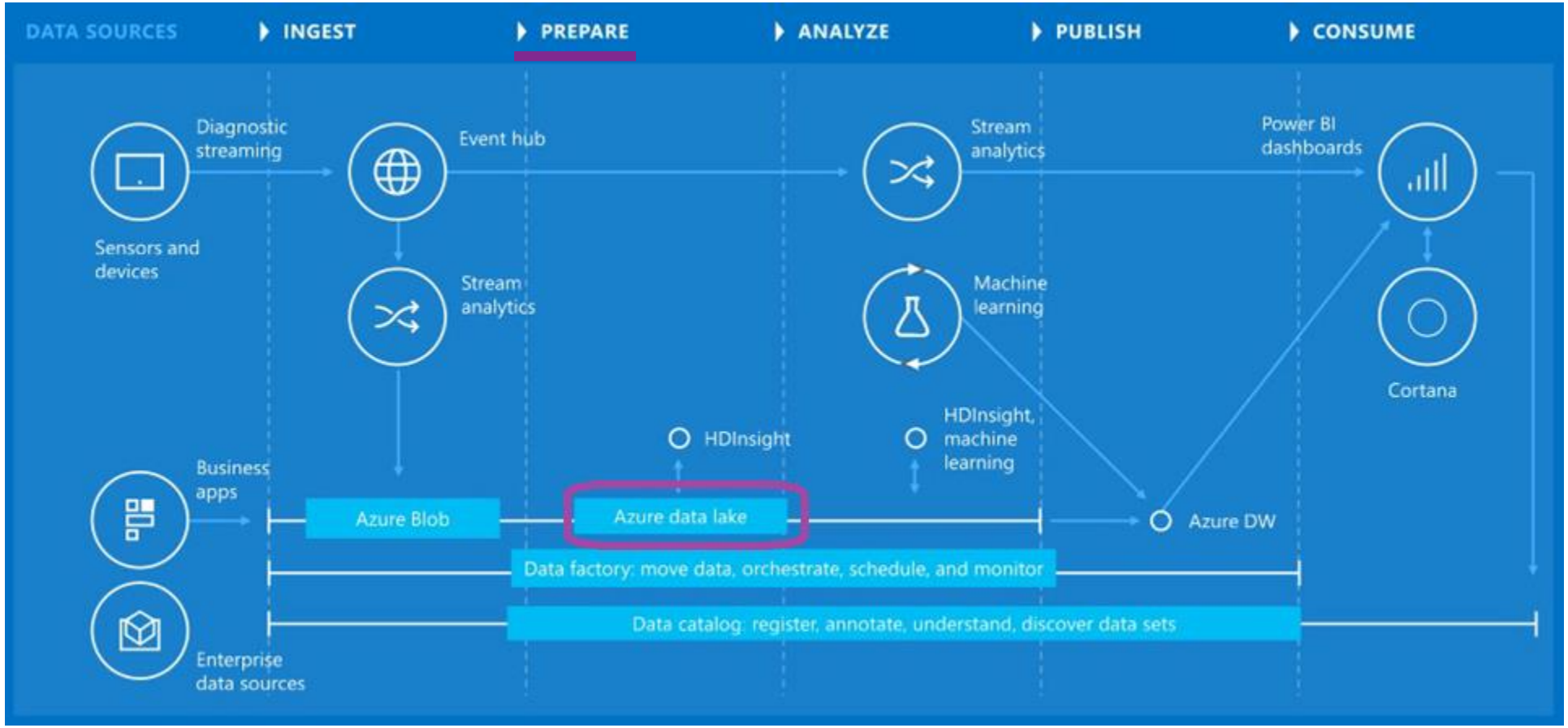
The Microsoft version…



https://azure.microsoft.com/en-gb/solutions/data-lake/

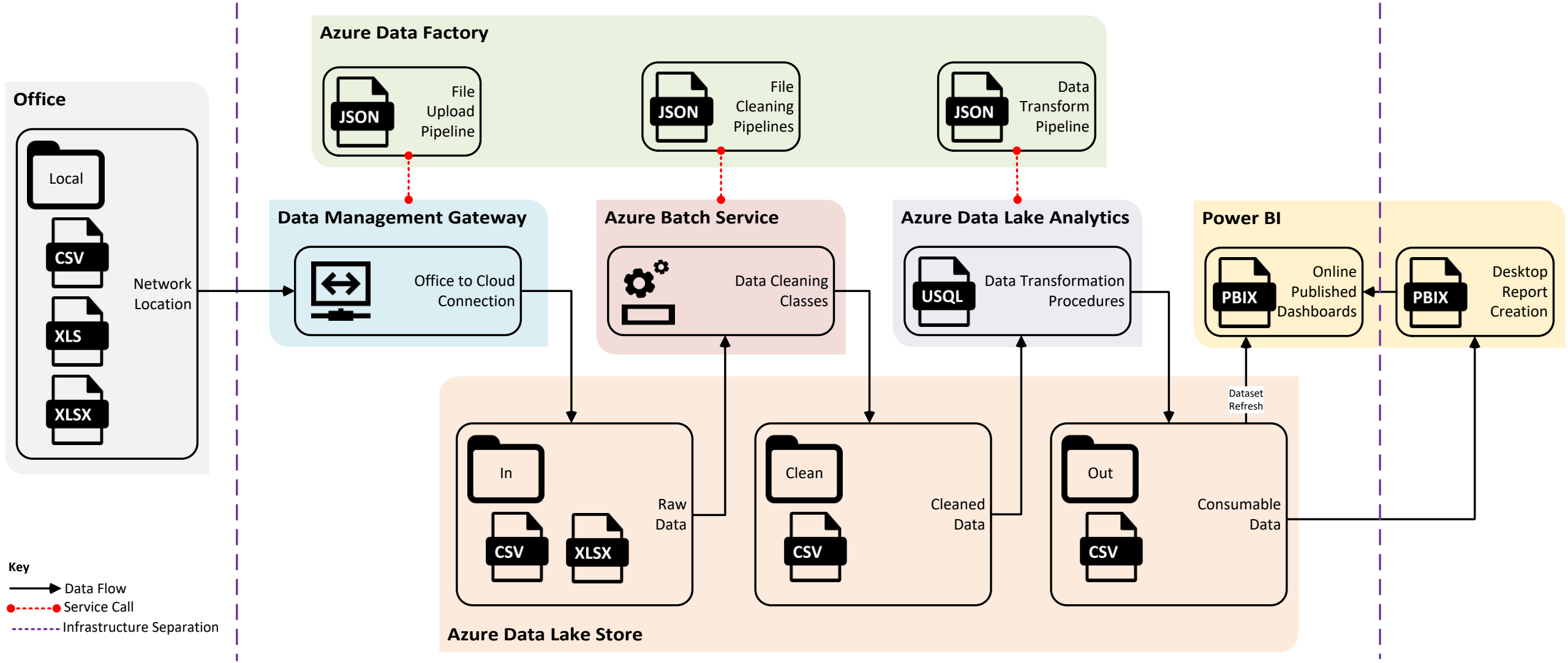
Azure Data Lake within the Cortana Intelligence Data Flow

- How does a lake fit into our data platform architecture?
- Is Data Lake going to run in isolation or be part of a larger pipeline?



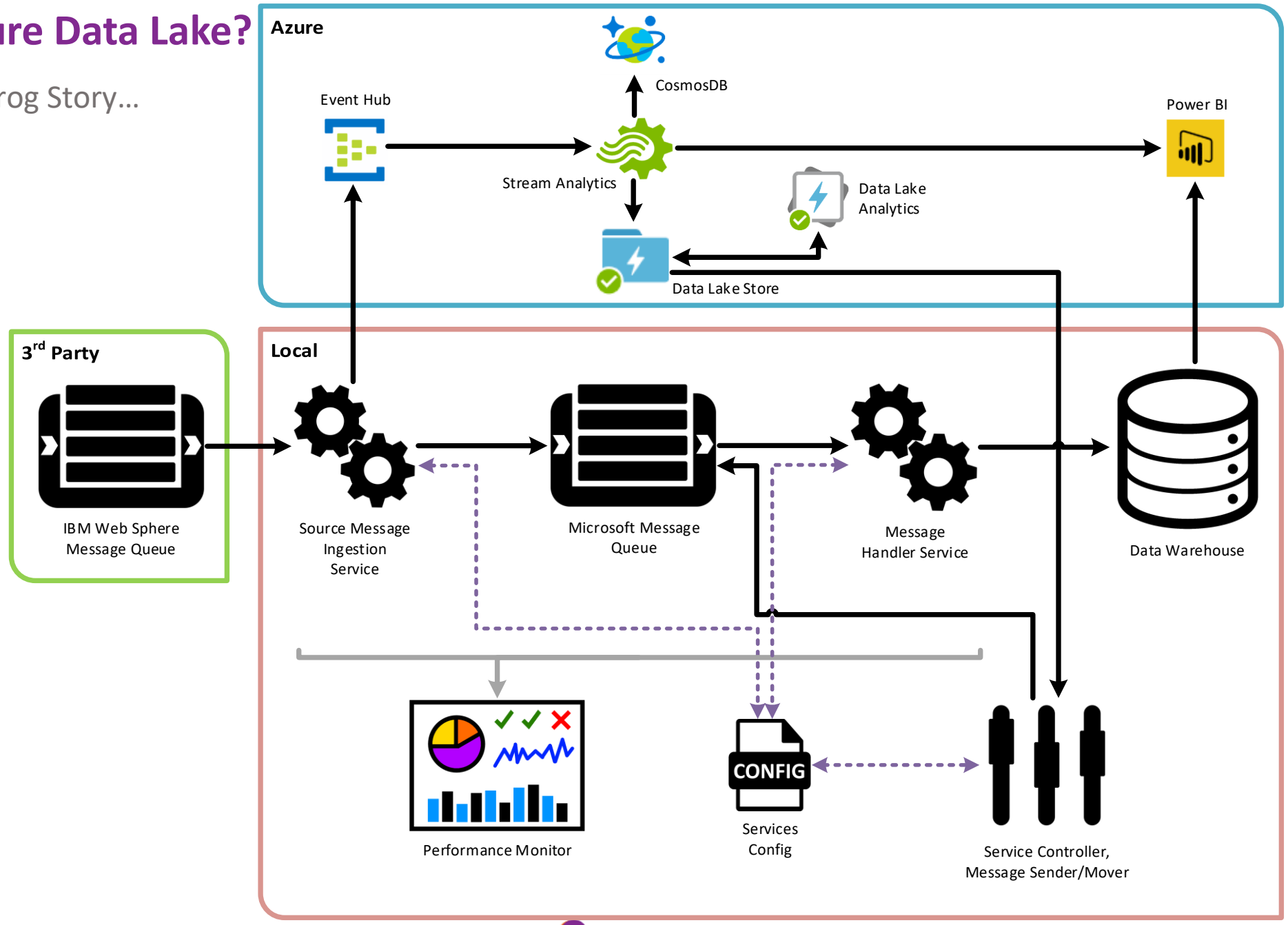
Why Use Azure Data Lake?

A Purple Frog Story…



Why Use Azure Data Lake?

Another Purple Frog Story…



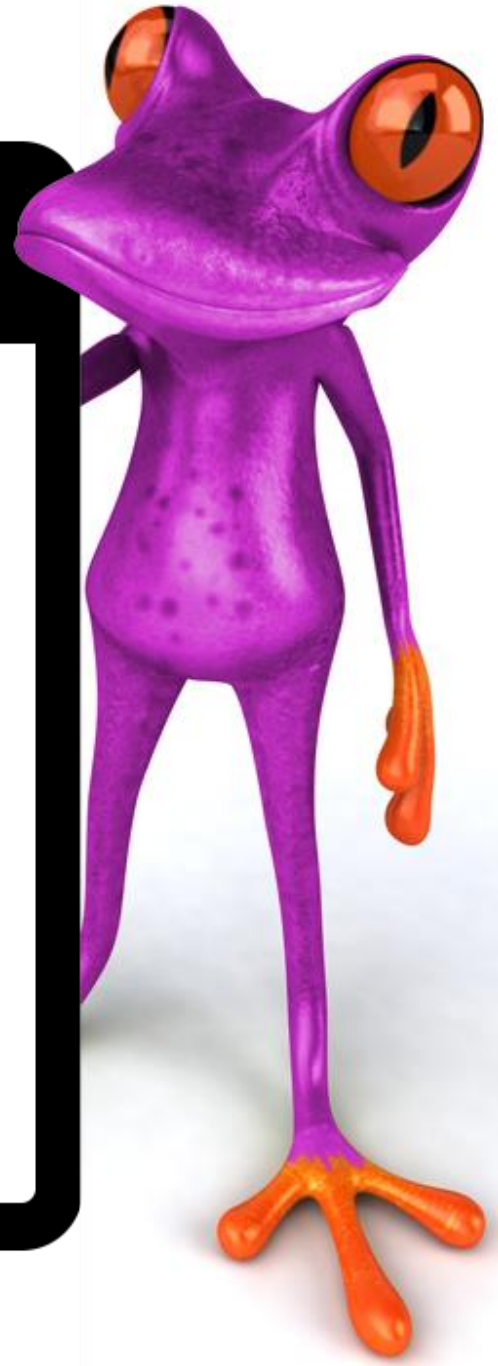
Session Agenda

What is Azure Data Lake?
- Storage & compute

Why use it?
- Real world examples

Working with each service

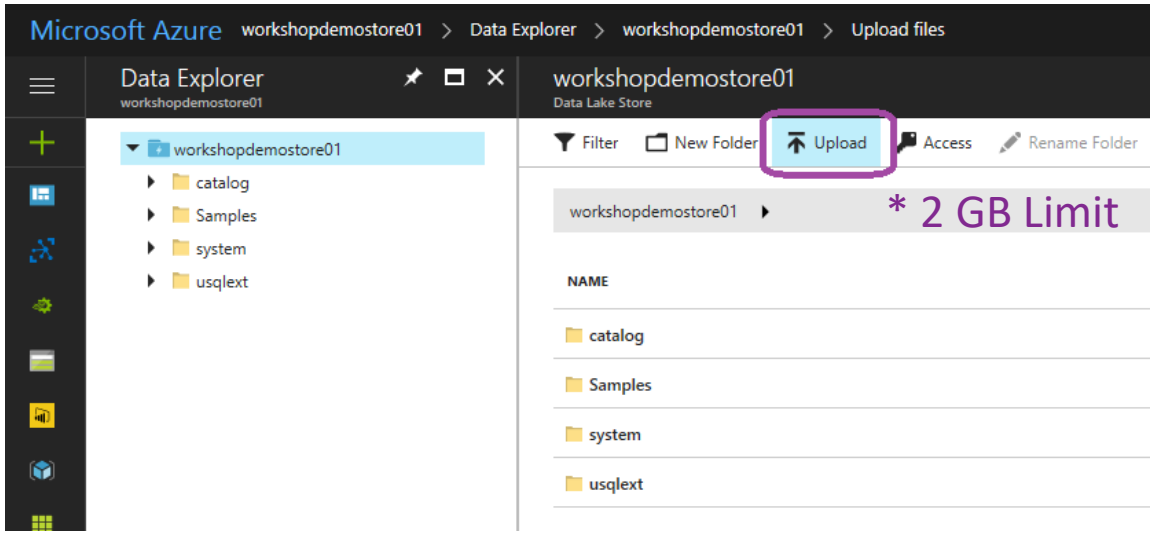
U-SQL
- The language (T-SQL, C#)
- Database objects
- Advanced capabilities



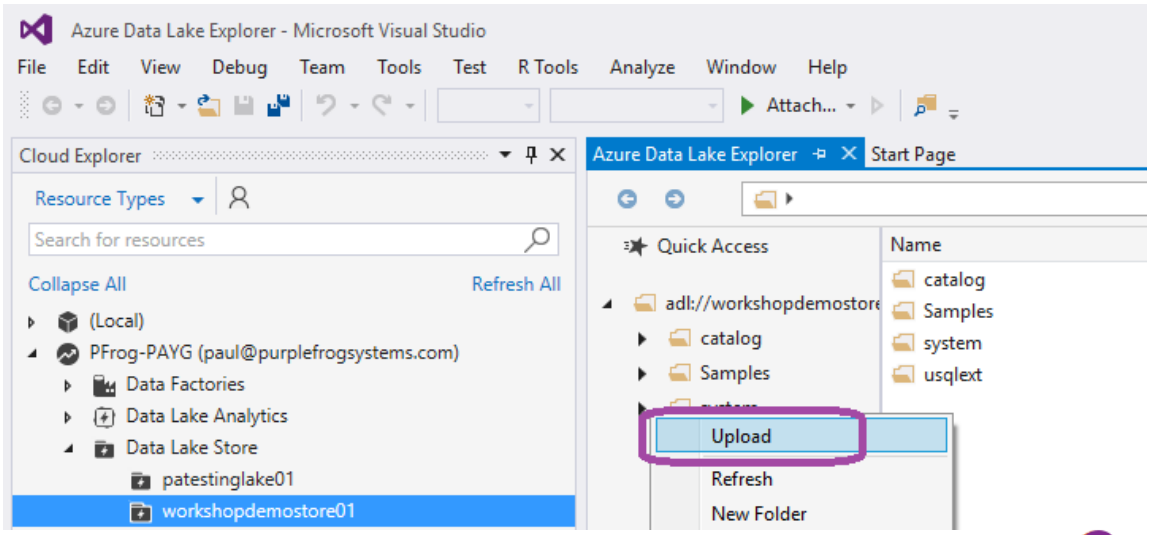
Uploading Data Into Azure Data Lake Store

Manual

Azure Portal UI



Visual Studio



Automatic

Azure Data Factory

```
{
"name": "DataLakeUploads",
  "properties": {
    "activities": [
      {"type": "Copy"},
      "inputs": [
          {"name": "OnPremFileServer"}
        ],
        "outputs": [
          {"name": "DataLakeStore"}
//etc...
```

Power Shell

```
Import-AzureRmDataLakeStoreItem `
    -AccountName $dataLakeStoreName `
    -Path "C:\Temp\SampleData.txt"
    -Destination $myrootdir\Samples\Data.txt
```

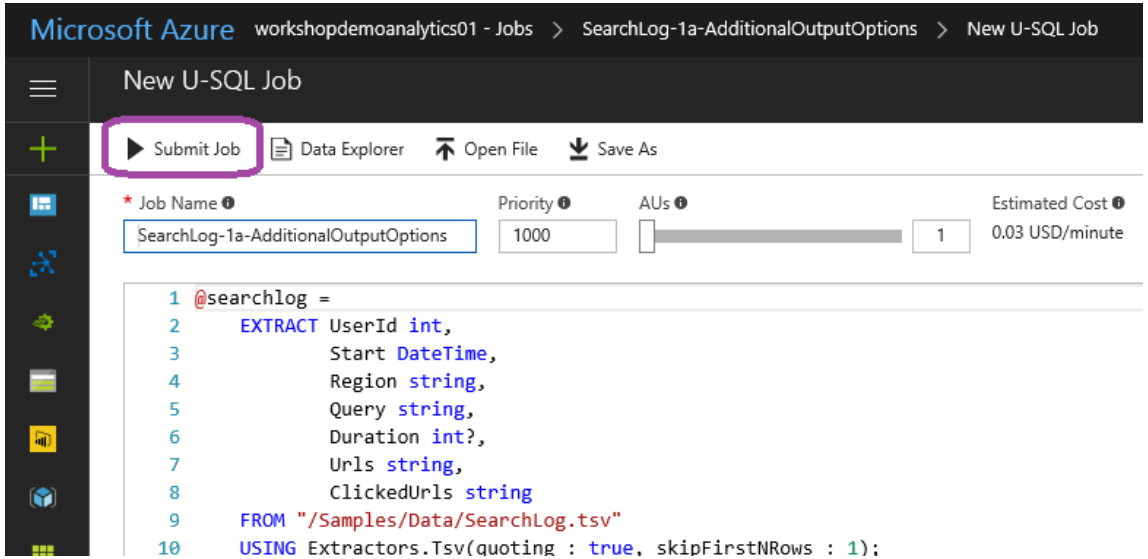
.Net SDK

```
var parameters = new UploadParameters(
                                        srcFilePath,
                                        destFilePath,
                                        accName, isOverwrite: force);

var frontend = new DataLakeStoreFrontEndAdapter(
                                      accName, adlsFileSstemClient);

var uploader = new DataLakeStoreUploader(parameters,
                                               frontend);
uploader.Execute();
```

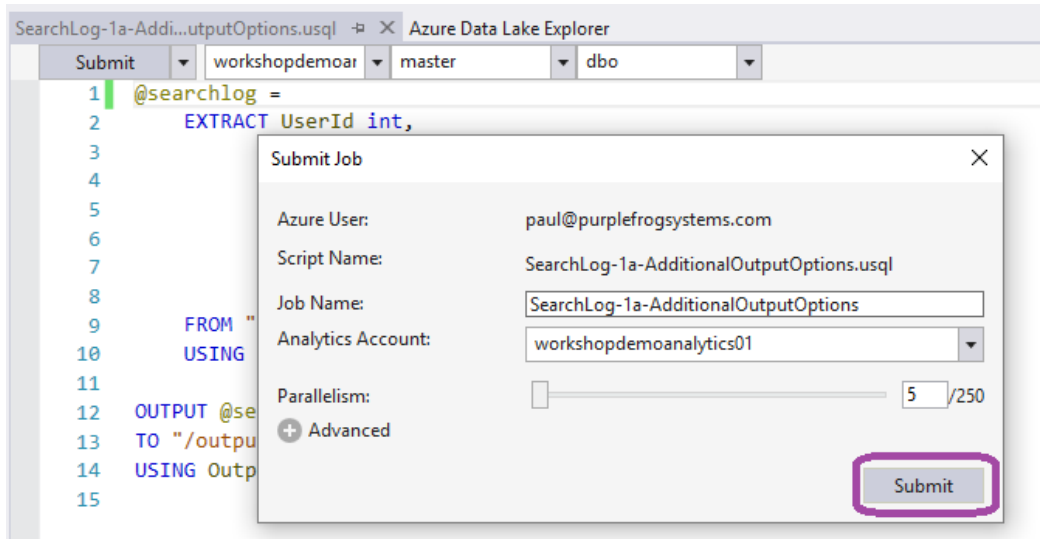
Triggering Azure Data Lake Analytics Jobs

Manual

Azure Portal UI



Visual Studio



Automatic

Azure Data Factory

```
{
"name": "DataLakeProcessing",
  "properties": {
    "activities": [
      {"type": "DataLakeAnalyticsU-SQL"},
       "typeProperties": {
                  "script": "@Data = SELECT * FROM @CSVFile",
           //OR   "script": "[dbo].[usp_ProcessData](@TimeSlice)",
           //OR   "scriptPath": "BlobStore\ProcessData.USQL",
                  "scriptLinkedService": "BlobStore",
                  "degreeOfParallelism": 2,
                  "priority": 7
```

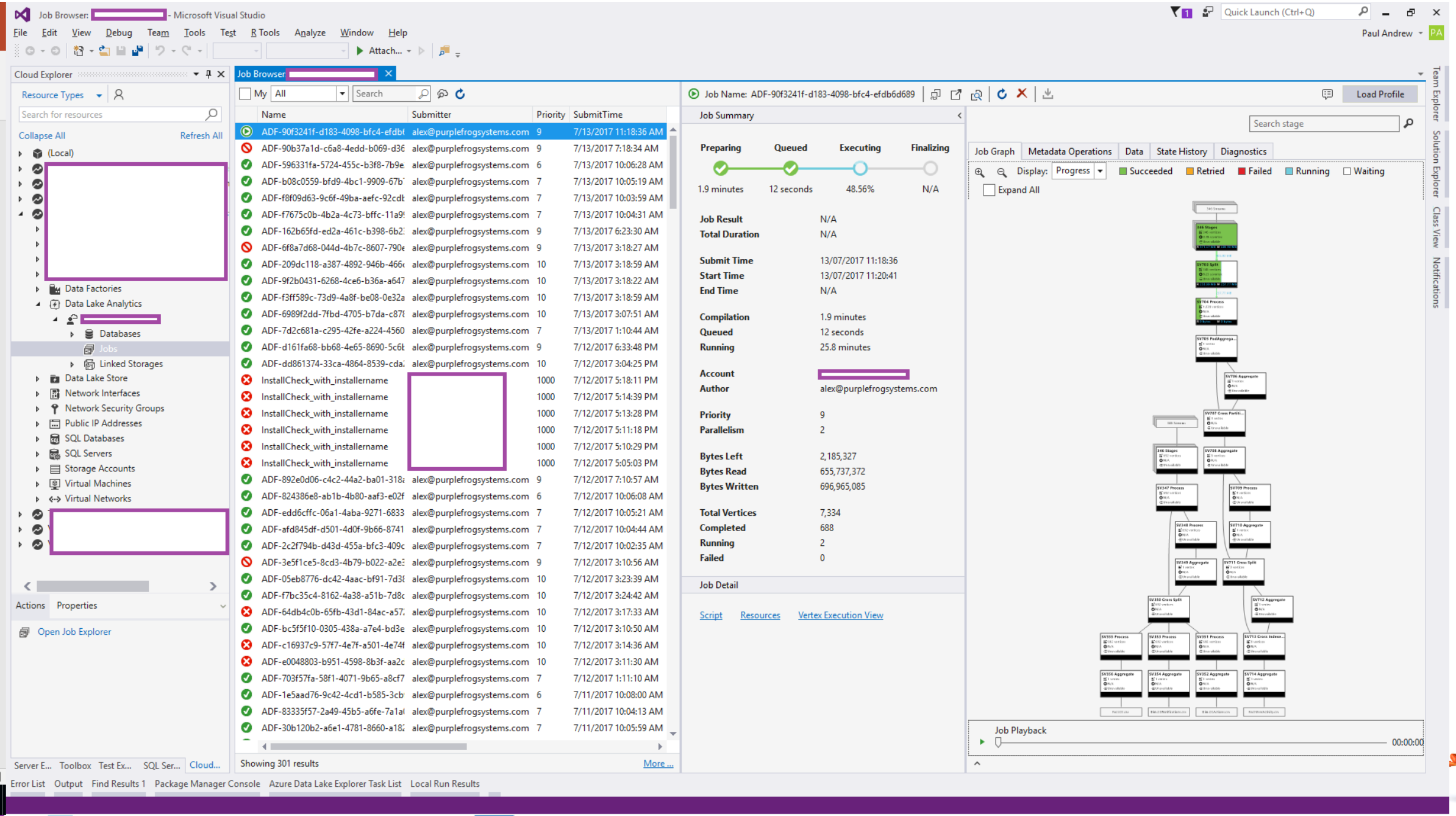
Power Shell

```
Submit-AzureRmDataLakeAnalyticsJob `
    -Name "ProcessData" `
    -AccountName $dataLakeAnalyticsName `
    -ScriptPath $usqlScript
```

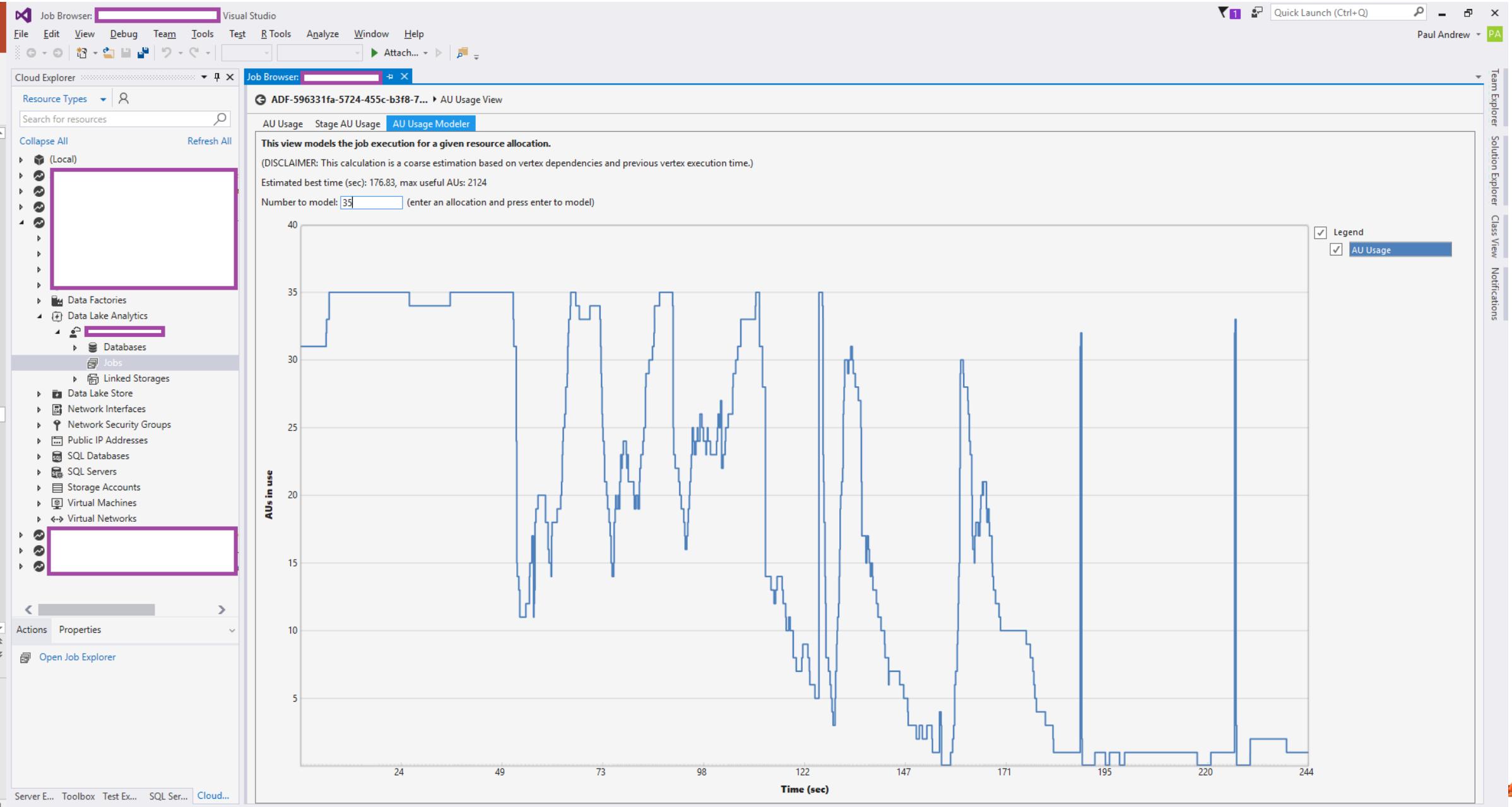
.Net SDK

```
var script = File.ReadAllText(scriptPath);
var jobId = Guid.NewGuid();
var properties = new USqlJobProperties(script);
var parameters = new JobInformation(jobName, JobType.USql,
           properties, priority: 1, degreeOfParallelism: 1,
           jobId: jobId);

var jobInfo = _adlaJobClient.Job.Create(_adlaAccountNamc
                                        jobId, parameters)

return jobId;
```

Job Management – Data Flow Graphs



Job Management – AU Modelling



Job Management – AU Polices & Usage Costs

Microsoft Azure

Data Lake Analytics

Save Discard

Name	Status	Location	Created
	Running	eastus2	09/01/2017 09:33:34 AM

Subscription Name

Piriform Pay-As-You-Go

Change subscription

Subscription Id

Debug Data Access Levels

Who has permission to view debug vertex data in Visual Studio

Users of this account and Microsoft Support Engineers

Maximum AUs

Set the maximum AUs for this account.

250 AU

Maximum Number of Running Jobs

Set the maximum number of jobs that you can run at the same time.

10 Jobs

Days to Retain Job Queries

Set the number of days to keep U-SQL job resources (scripts, DLLs, graphs) in your Data Lake Store account.

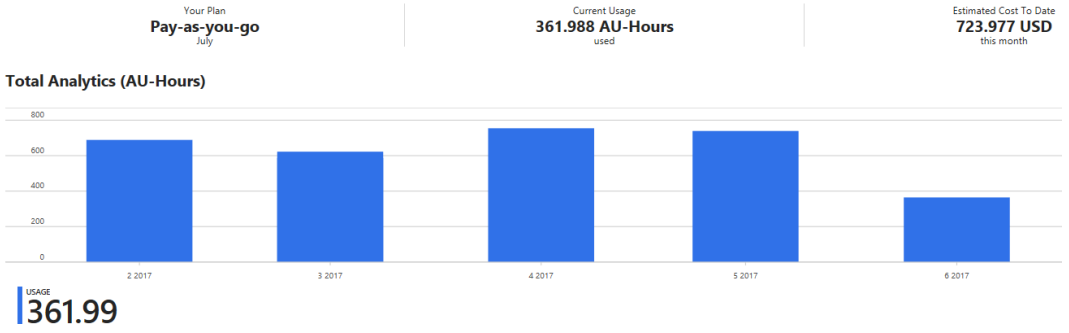
30 Days

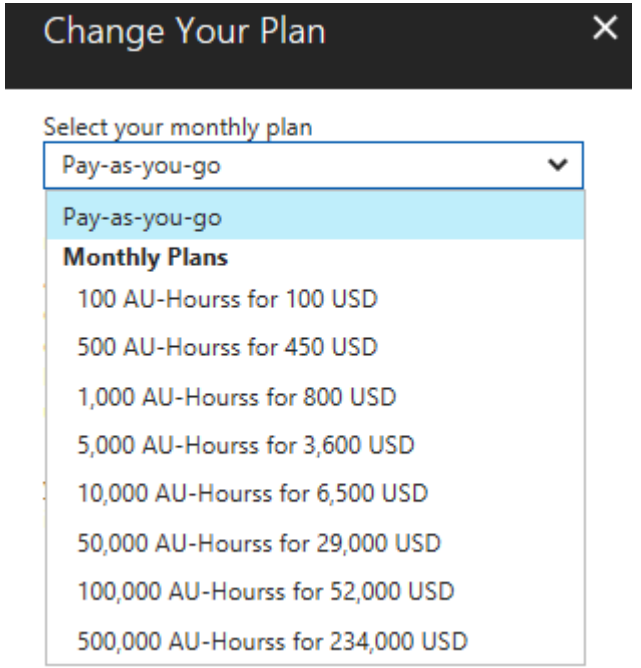
Job Submission Limits

Set the maximum amount of AUs and the highest priority each user has access to when they submit a job. You can override the default by adding policies for different users, and groups. If a submitter has multiple policies that apply to them, the most permissive limit will take effect.

Add Policy

Bulk buy your compute!



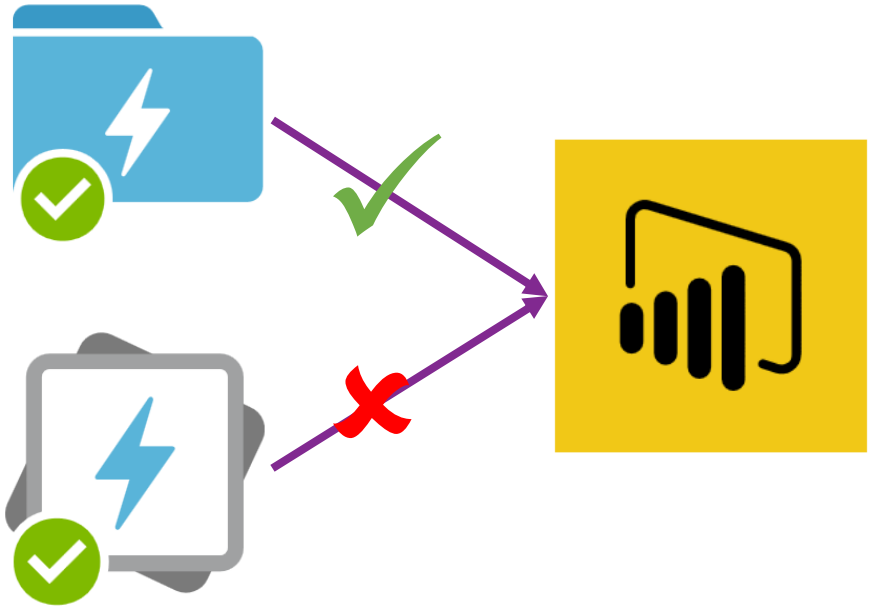


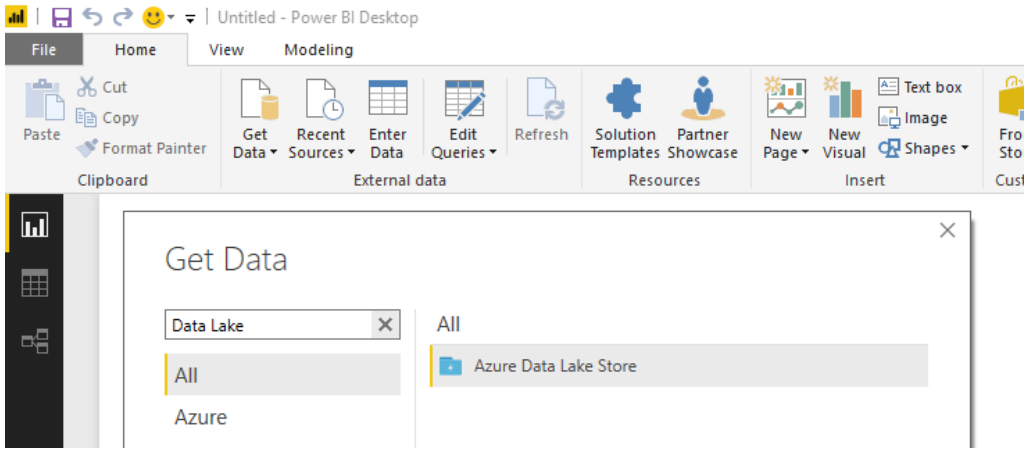
Pay-as-you-go
400 AU's
800 USD

Pay Monthly
500 AU's
450 USD

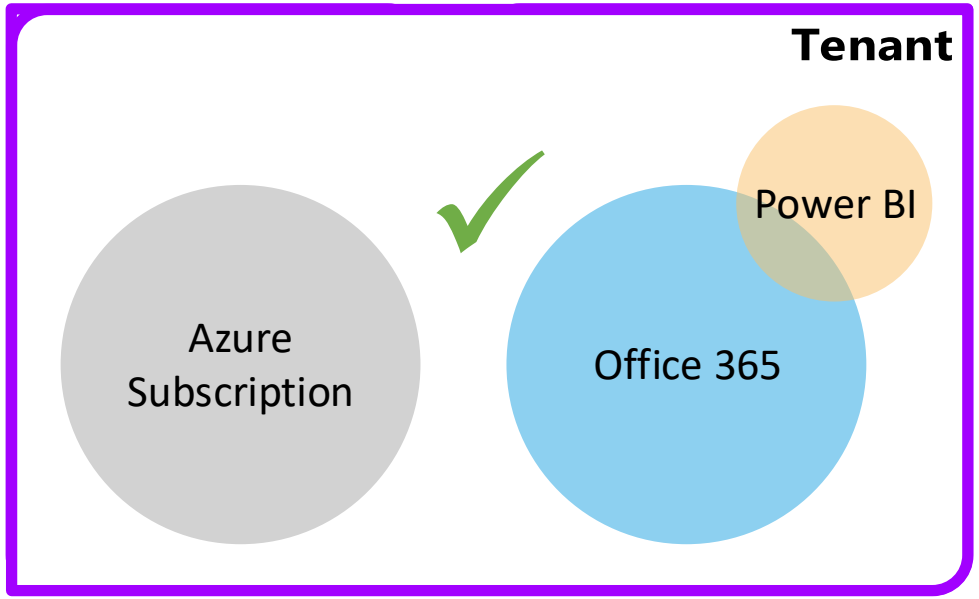
Consuming Data Lake with Power BI

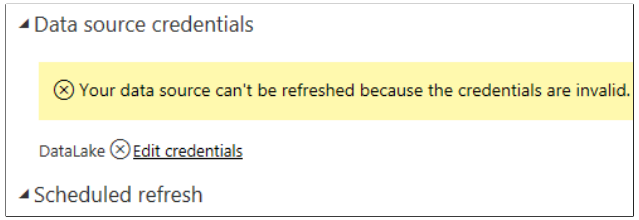
Only storage connector available at present.





Published PowerBI.com dataset refresh across tenants.





"Failed to update data source credentials:
The credentials provided for the DataLake source are invalid."

https://www.purplefrogsystems.com/paul/2017/06/connecting-power-bi-to-azure-data-lake-store-across-tenants/

Session Agenda

What is Azure Data Lake?

- Storage & compute

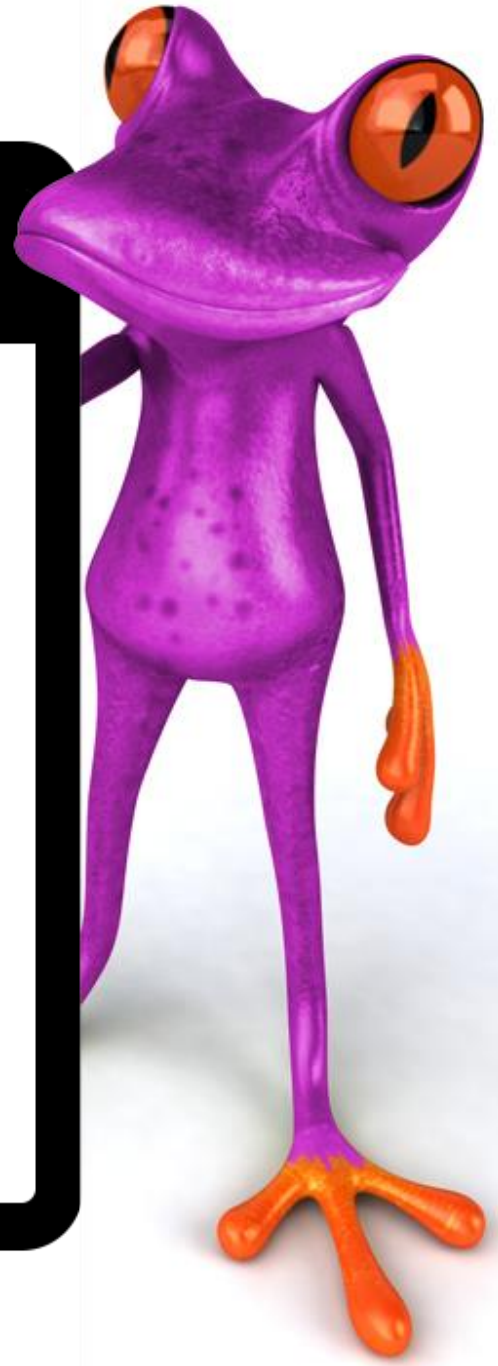
Why use it?

- Real world examples

Working with each service

U-SQL

- The language (T-SQL, C#)
- Database objects
- Advanced capabilities



What is U-SQL?

- Hybrid of T-SQL and C#
- Created by Michael Rys

```
@FileSummary= SELECT CountryCode,
              COUNT(*) AS RecordCount,
              SUM(SMSIn) AS SMSIn,
              SUM(SMSOut) AS SMSOut,
              SUM(CallIn) AS CallIn,
              SUM(CallOut) AS CallOut,
              SUM(Data) AS Data
        FROM @RawData
        GROUP BY CountryCode;

@DataPrep01 =
    SELECT *
           , DateTime.Parse("1970-01-01").AddSeconds(TimeMS/1000) AS StartWindowDT
           , DateTime.Parse("1970-01-01").AddSeconds(TimeMS/1000).AddMinutes(10) AS EndWindowDT
    FROM @RawData;
```

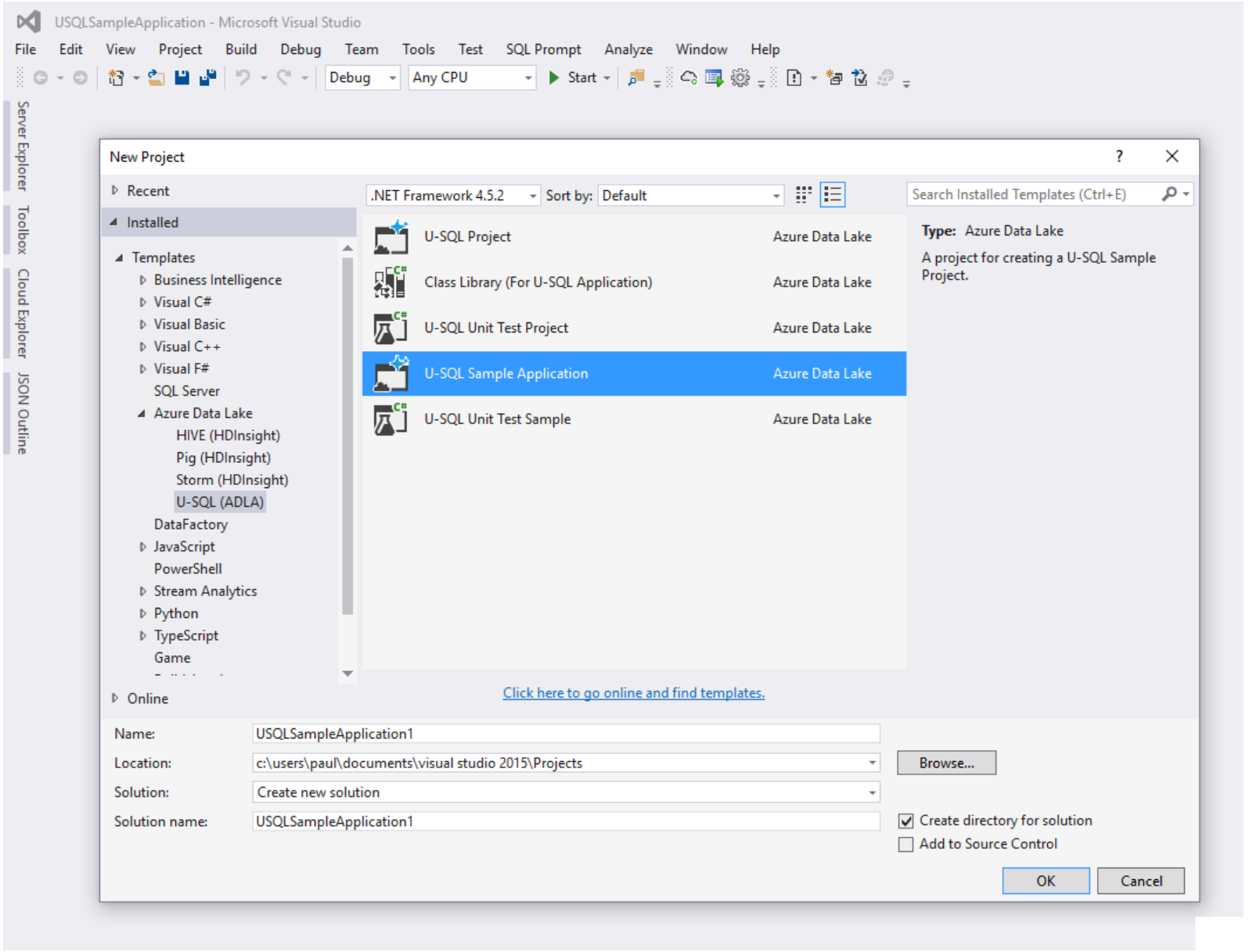
@MikeDoesBigData
#USQL

Why 'U'?

- Unifies T-SQL and C#
- Dive into the data lake like a U-Boat.
- U is the next letter after T-SQL.

Developing U-SQL – The Tools

- Visual Studio 2013
  - Azure SDK – ADL Tools
- Visual Studio 2015
  - Azure SDK – ADL Tools
- Visual Studio 2017
  - Out of the box extensions
- Azure Portal
  - Directly with blades
- Visual Studio Code
  - Extension now GA



U-SQL, like T-SQL

- Based on T-SQL
  - SELECT
  - DISTINCT
  - SUM()
  - FROM
  - GROUP BY
  - ORDER BY
  - Windowing Functions – ROW_NUMBER() etc.

```
@FileSummary= SELECT CountryCode,
                     COUNT(*) AS RecordCount,
                     SUM(SMSIn) AS SMSIn,
                     SUM(SMSOut) AS SMSOut,
                     SUM(CallIn) AS CallIn,
                     SUM(CallOut) AS CallOut,
                     SUM(Data) AS Data
              FROM @RawData
              GROUP BY CountryCode;
```

U-SQL, but with C#

- Uses .Net Data Types

bool	decimal	int	short	ushort
byte	double	long	uint	string
char	float	sbyte	ulong	

- Embeds C# functions

```
DateTime.Parse("1970-01-01").AddSeconds(TimeMS/1000) AS
```

- Access C# Assemblies

```
USQLApplication2.Udfs.ValidateDate(CustomerDoB) AS CustomerDoB
```


T-SQL vs U-SQL

T-SQL	U-SQL
CASE WHEN Gender IN (‘M’, ‘F’) THEN Gender ELSE ‘-’ END	(Gender=="M" || Gender=="F") ? Gender : "-" AS Gender
SubString(Field,0,10)	Field.SubString(0,10)
CharIndex(‘?’,Field)	Field.IndexOf(“?”)
IsNull(Field,’Unknown’)	String.IsEmptyOrNull(Field) ? Field : “Unknown”
Convert(Varchar, OrderDateTime, 112)	OrderDateTime.ToString(“yyyyMMdd”)
Replace(Field,’?’,’Unknown’)	Field.Replace(“?”,”Unknown”)

A Couple of Limitations

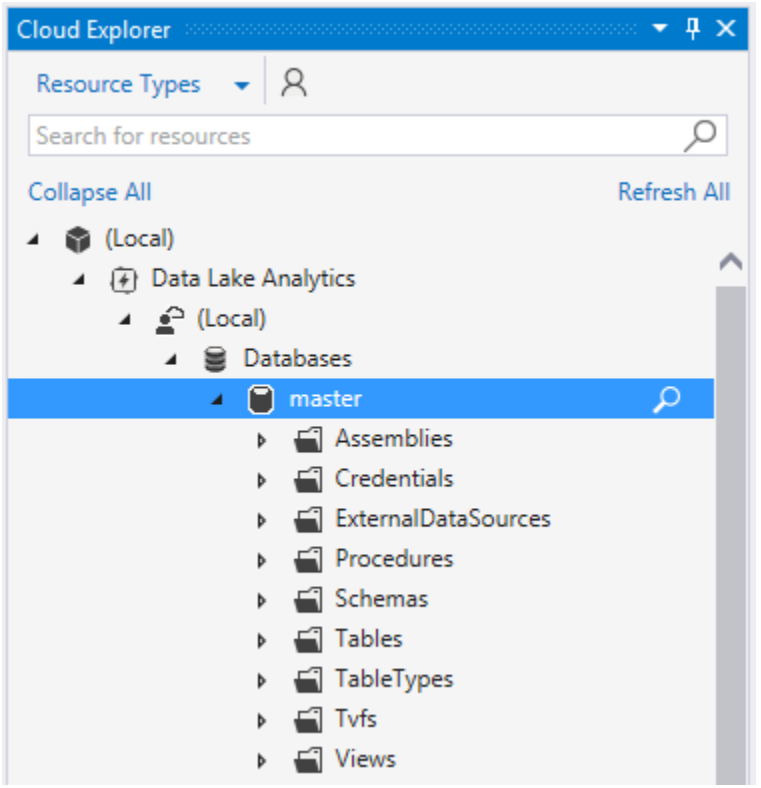
T-SQL	U-SQL
MERGE/UPDATE	https://www.purplefrogsystems.com/paul/2016/12/writing-a-u-sql-merge-statement/
WHILE	https://www.purplefrogsystems.com/paul/2017/05/recursive-u-sql-with-powershell-u-sql-looping/
Date Dimension	https://www.purplefrogsystems.com/paul/2017/02/creating-a-u-sql-date-dimension-numbers-table-in-azure-data-lake/

```
SELECT
        @Check = COUNT(0)
 FROM
        [SomeWhere]

IF @Check > 5
```

Azure Data Lake Analytics Database Object (DDL and DML)

- Assemblies
- Stored Procedures
- Schemas
- Tables
- Functions

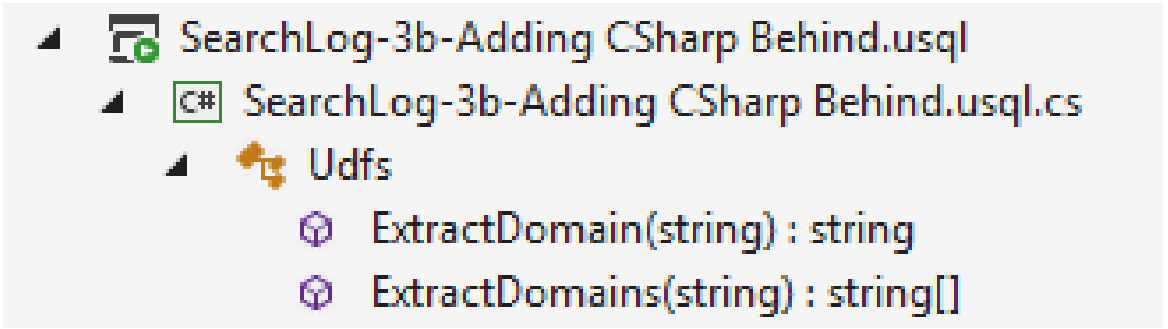
- Live in the '_catalog_' directory of your storage.




```
CREATE DATABASE IF NOT EXISTS [Demo];

CREATE SCHEMA IF NOT EXISTS [fact];

DROP TABLE IF EXISTS [fact].[Sales];

CREATE TABLE IF NOT EXISTS [fact].[Sales];

DROP PROCEDURE IF EXISTS [Demo].[fact].[usp_LoadSales];

CREATE PROCEDURE [Demo].[fact].[usp_LoadSales](@TimeSliceStart string)
AS
BEGIN /* do stuff */ END;
```

C# Code Behind

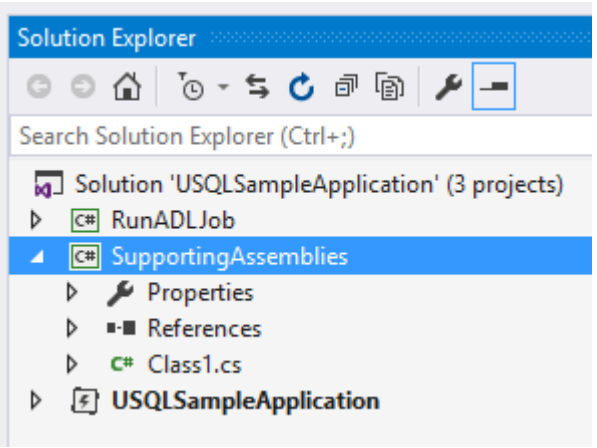
- Handled by Visual Studio when summiting jobs.
- U-SQL file does not carry C# around with it.
- Calling from other services it needs help compiling the C#.
- Calling U-SQL Stored procedures needs help.
- Assemblies are your friend ☺




```
CREATE ASSEMBLY IF NOT EXISTS [Demo].[USQLHelperClasses]
FROM "Assemblies/USQLHelperClasses.dll";
```

Complied C# Code in Assemblies

- Inline binary in U-SQL.
- Create the assembly from Azure Data Lake Store.
- Create the assembly from Azure Blob Storage.



https://www.purplefrogsystems.com/paul/2017/04/calling-u-sql-stored-procedures-with-c-code-behind/

Demo Time!

U-SQL Further Reading

Microsoft Blog – An Introduction to U-SQL in Azure Data Lake
https://blogs.msdn.microsoft.com/robinlester/2016/01/04/an-introduction-to-u-sql-in-azure-data-lake/

Microsoft Documentation – U-SQL Programmability Guide
https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-u-sql-programmability-guide

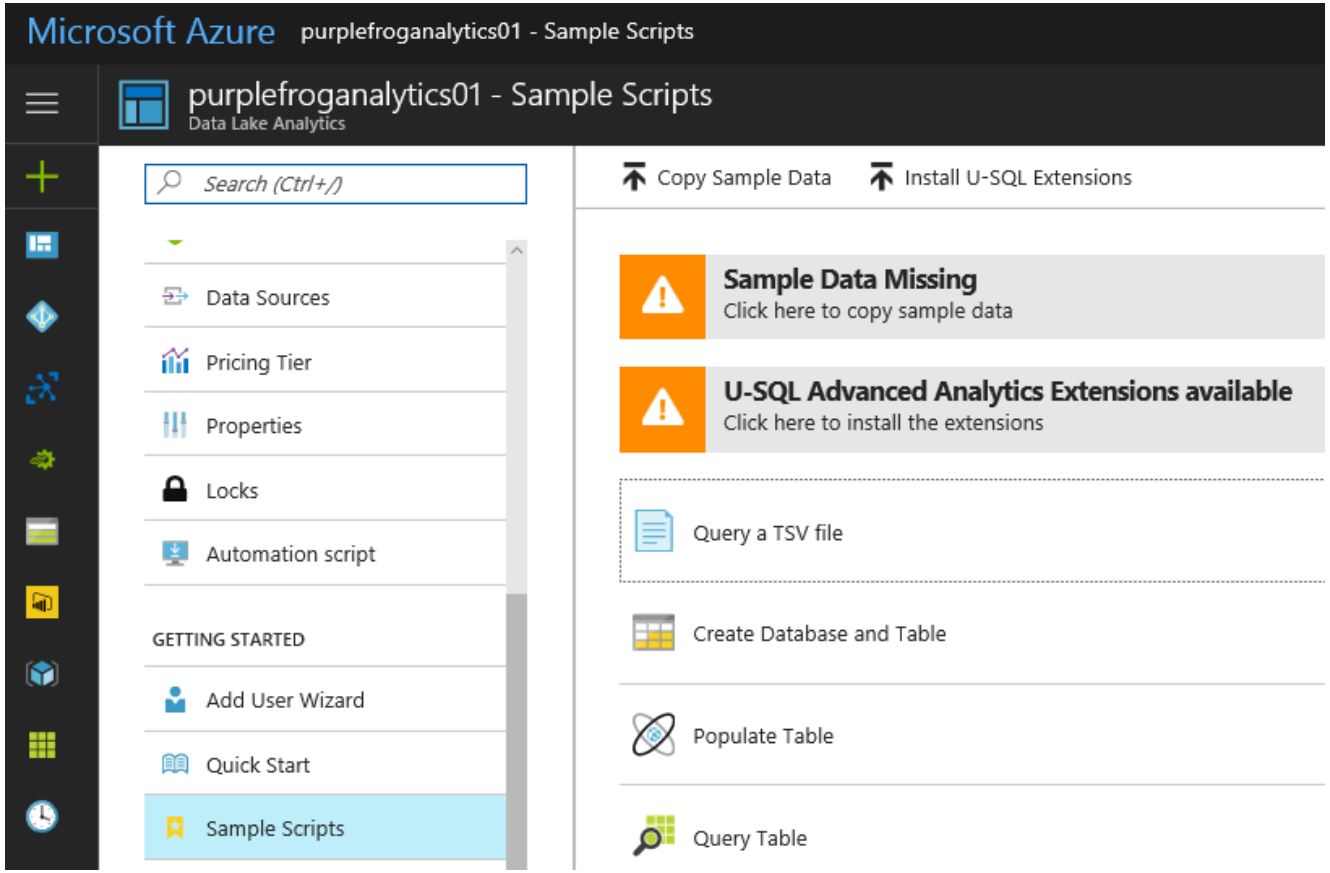
Microsoft MSDN – U-SQL Language Reference
https://msdn.microsoft.com/en-US/library/azure/mt591959(Azure.100).aspx

SQL Server Central – Stairway to U-SQL
http://www.sqlservercentral.com/stairway/142480/

Purple Frog Blog
https://www.purplefrogsystems.com/

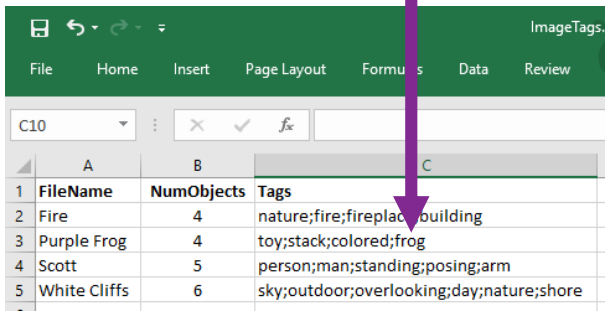
Stack Overflow – U-SQL Tag
http://stackoverflow.com/questions/tagged/u-sql

Using Cognitive Services with Azure Data Lake



https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-u-sql-cognitive

```
USE [CognitiveServices];

REFERENCE ASSEMBLY ImageCommon;
REFERENCE ASSEMBLY FaceSdk;
REFERENCE ASSEMBLY ImageEmotion;
REFERENCE ASSEMBLY ImageTagging;
REFERENCE ASSEMBLY ImageOcr;

@imgs =
    EXTRACT
        [FileName] string,
        [ImgData] byte[]
    FROM
        @"/Images/{FileName}.jpg"
    USING
        new Cognition.Vision.ImageExtractor();

@imgTags =
    PROCESS
        @imgs
    PRODUCE
        [FileName],
        [NumObjects] int,
        [Tags] string
    READONLY
        [FileName]
    USING
        new Cognition.Vision.ImageTagger();

OUTPUT @imgTags
TO "/output/ImageTags.csv"
USING Outputters.Csv(quoting : true, outputHeader : tru
```



	A	B	C
1	FileName	NumObjects	Tags
2	Fire	4	nature;fire;fireplac building
3	Purple Frog	4	toy;stack;colored;frog
4	Scott	5	person;man;standing;posing;arm
5	White Cliffs	6	sky;outdoor;overlooking;day;nature;shore

Thank You