Azure Data Lake

The Services. The U-SQL. The C#.

Paul Andrew | Senior Data Analytics Consultant & Data Platform MVP

28 March 2018

MVP Microsoft® Most Valuable Professional

@MrPaulAndrew

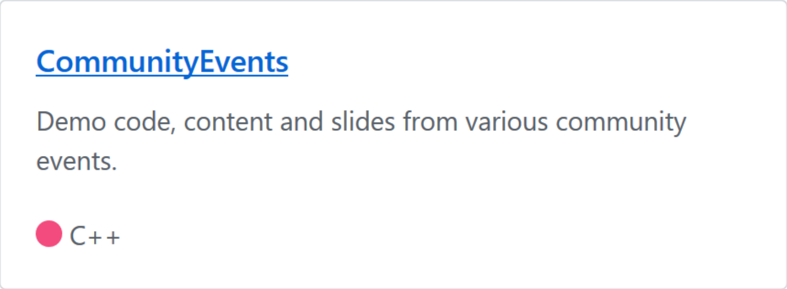
Microsoft Partner Microsoft

Gold Data Analytics
Gold Data Platform
Gold Cloud Platform

adatis

GitHub

https://github.com/mrpaulandrew



CommunityEvents

Demo code, content and slides from various community events.

● C++

Agenda

What is Azure Data Lake?

Storage & Compute

Why use Data Lake?

The Modern Data Warehouse

How can we work with Data Lake?

Development & Management

U-SQL

‘Hello World’ to Advanced Analytics

Agenda

What is Azure Data Lake?

Storage & Compute

Why use Data Lake?

The Modern Data Warehouse

How can we work with Data Lake?

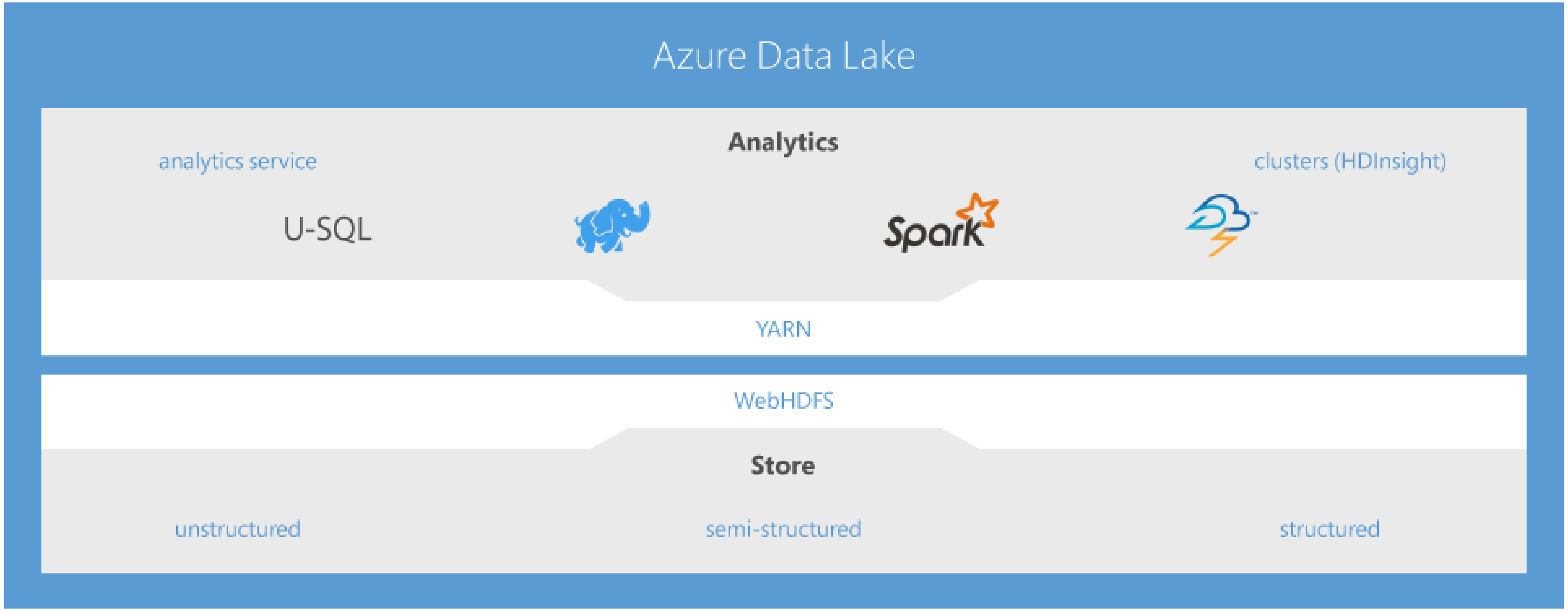
Development & Management

U-SQL

‘Hello World’ to Advanced Analytics

What is Azure Data Lake?

The Microsoft version:

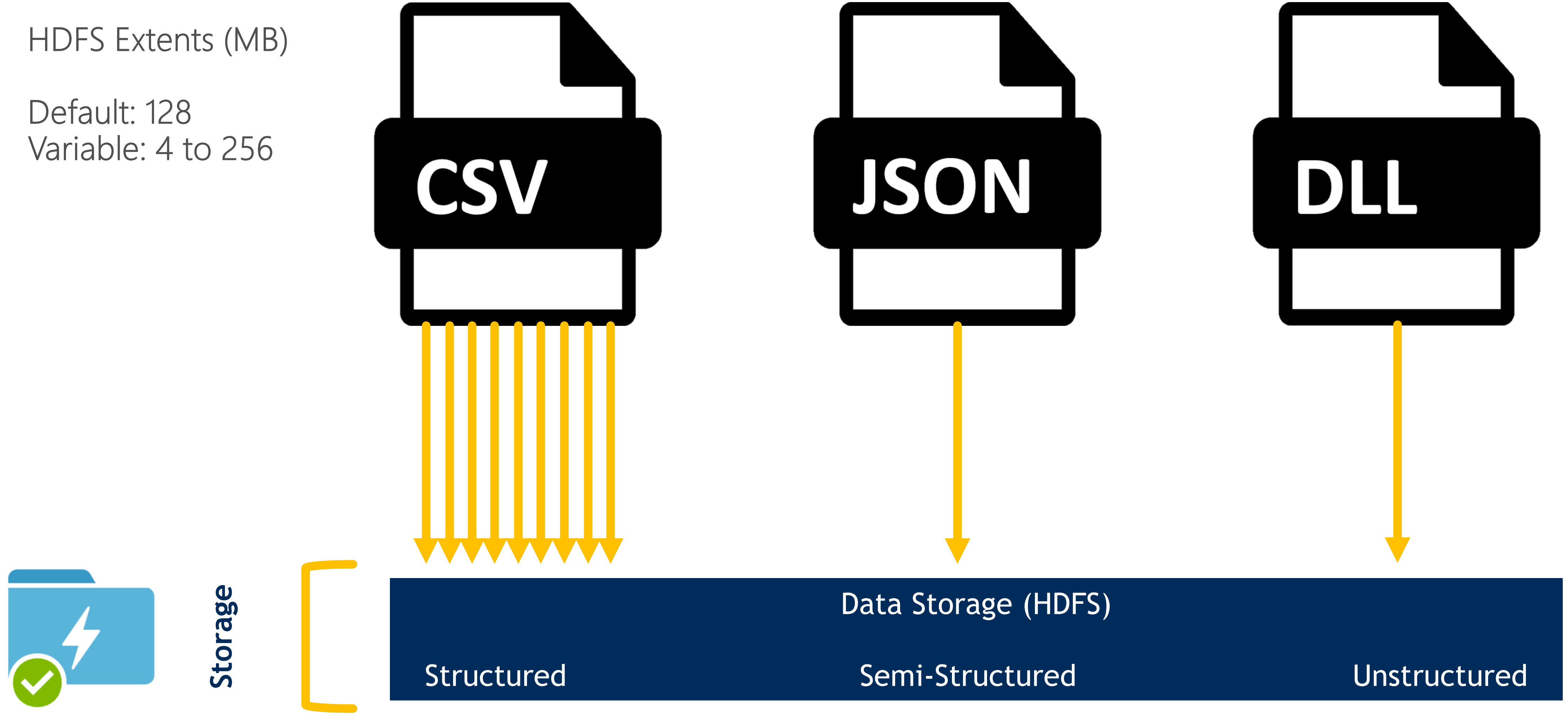


https://azure.microsoft.com/en-gb/solutions/data-lake/

What is Azure Data Lake?

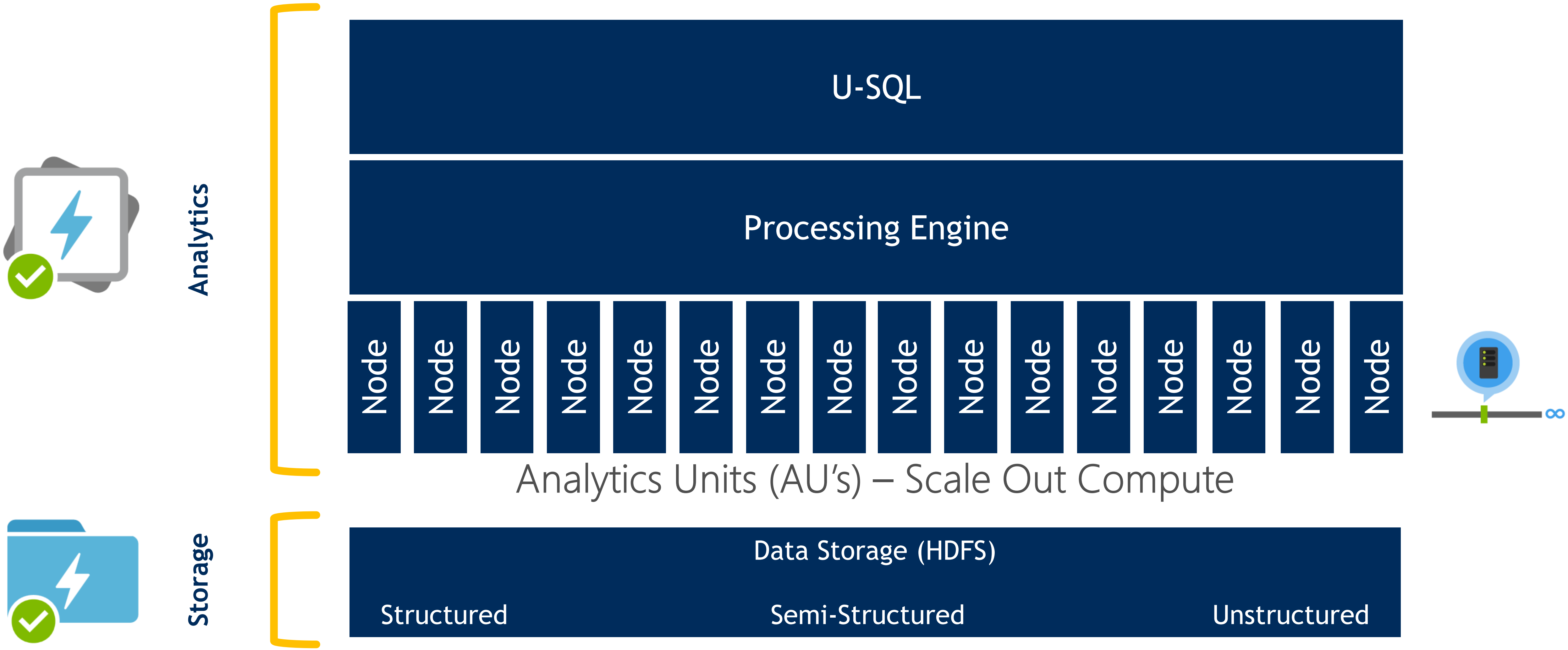
HDFS Extents (MB)

Default: 128
Variable: 4 to 256



https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-performance-tuning-guidance

What is Azure Data Lake?



Analytics

U-SQL

Processing Engine

Node Node Node Node Node Node Node Node Node Node Node Node Node Node Node Node

Analytics Units (AU's) – Scale Out Compute

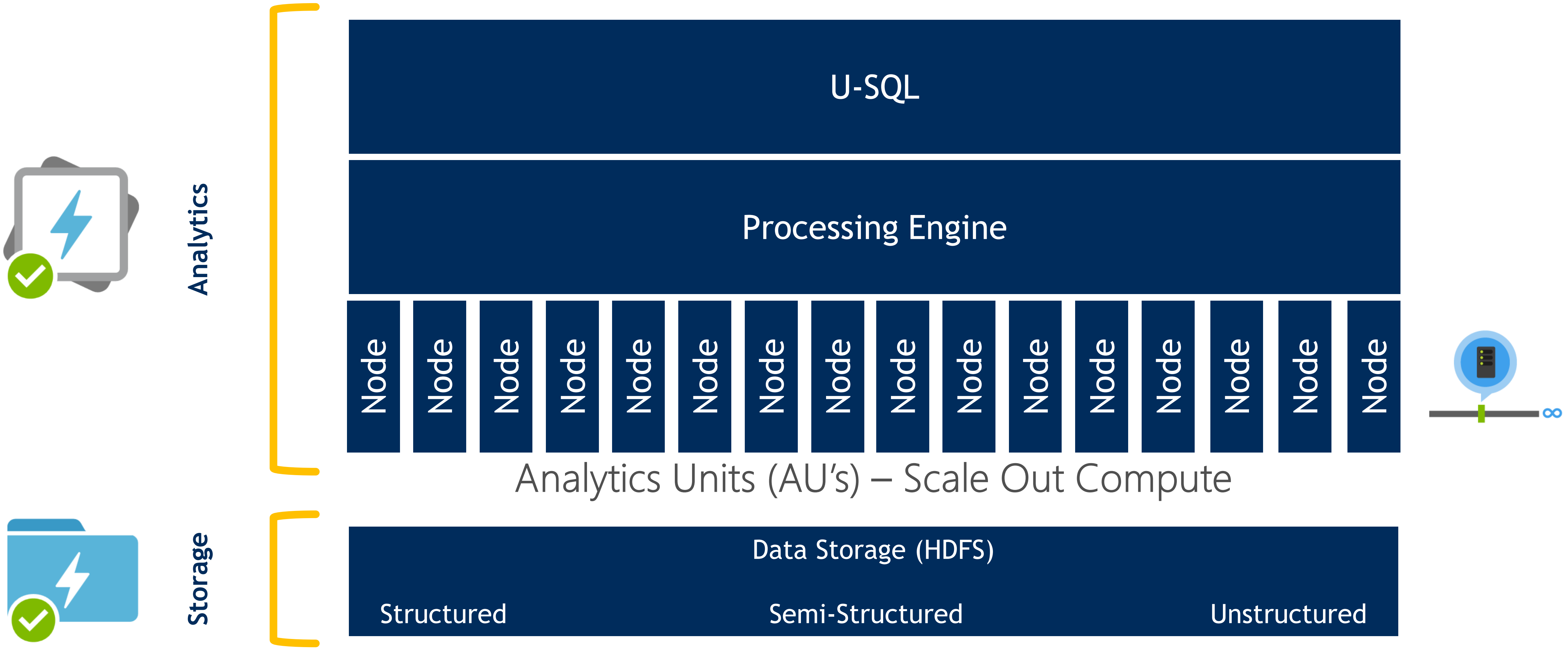
Storage

Data Storage (HDFS)

Structured Semi-Structured Unstructured

https://blogs.msdn.microsoft.com/azuredatalake/2016/10/12/understanding-adl-analytics-unit/

What is Azure Data Lake?



Analytics

U-SQL

Processing Engine

Node Node Node Node Node Node Node Node Node Node Node Node Node Node Node Node

∞

Analytics Units (AU's) – Scale Out Compute

Storage

Data Storage (HDFS)

Structured

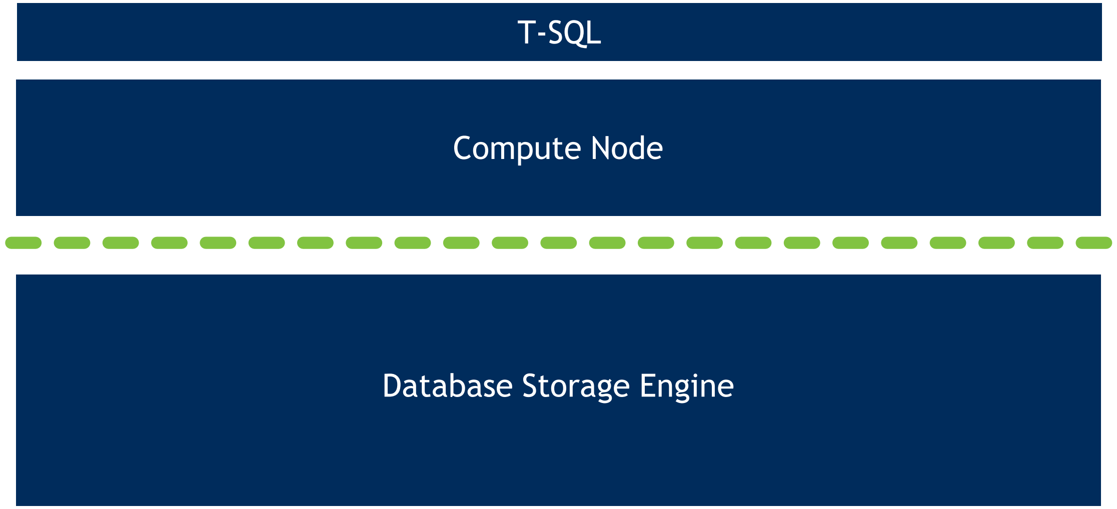
Semi-Structured

Unstructured

https://blogs.msdn.microsoft.com/azuredatalake/2016/10/12/understanding-adl-analytics-unit/

Azure Data Lake vs Other Data Services

Azure SQL Database (SQLDB)



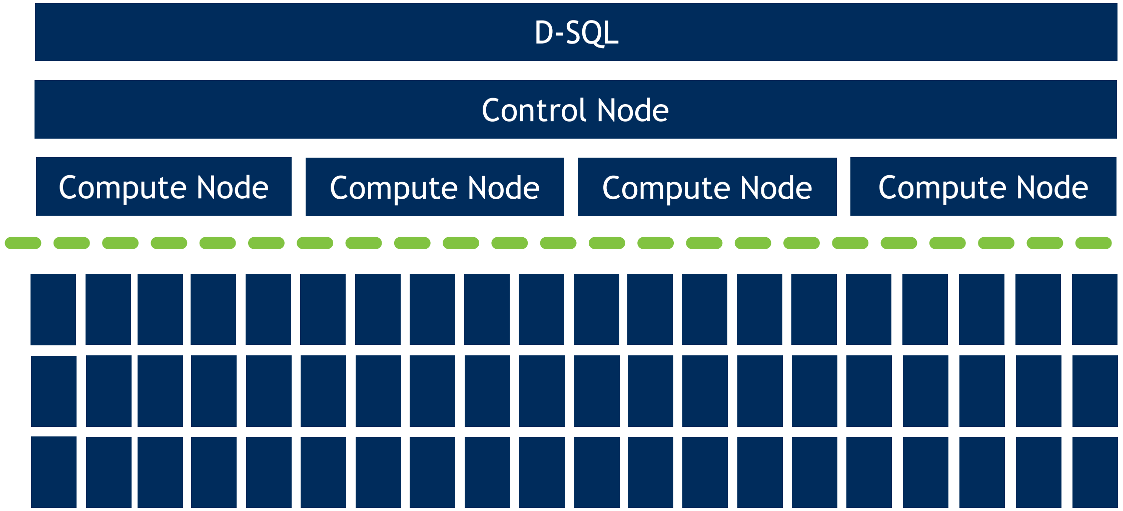
T-SQL

Compute Node

Database Storage Engine

Database Transaction Units (DTU's) – Scale Up Compute

Azure SQL Data Warehouse (SQLDW)



D-SQL

Control Node

Compute Node

Compute Node

Compute Node

Compute Node

Database Warehouse Units (DWU's) – Scale Out Compute

Agenda

What is Azure Data Lake?

Storage & Compute

Why use Data Lake?

The Modern Data Warehouse

How can we work with Data Lake?

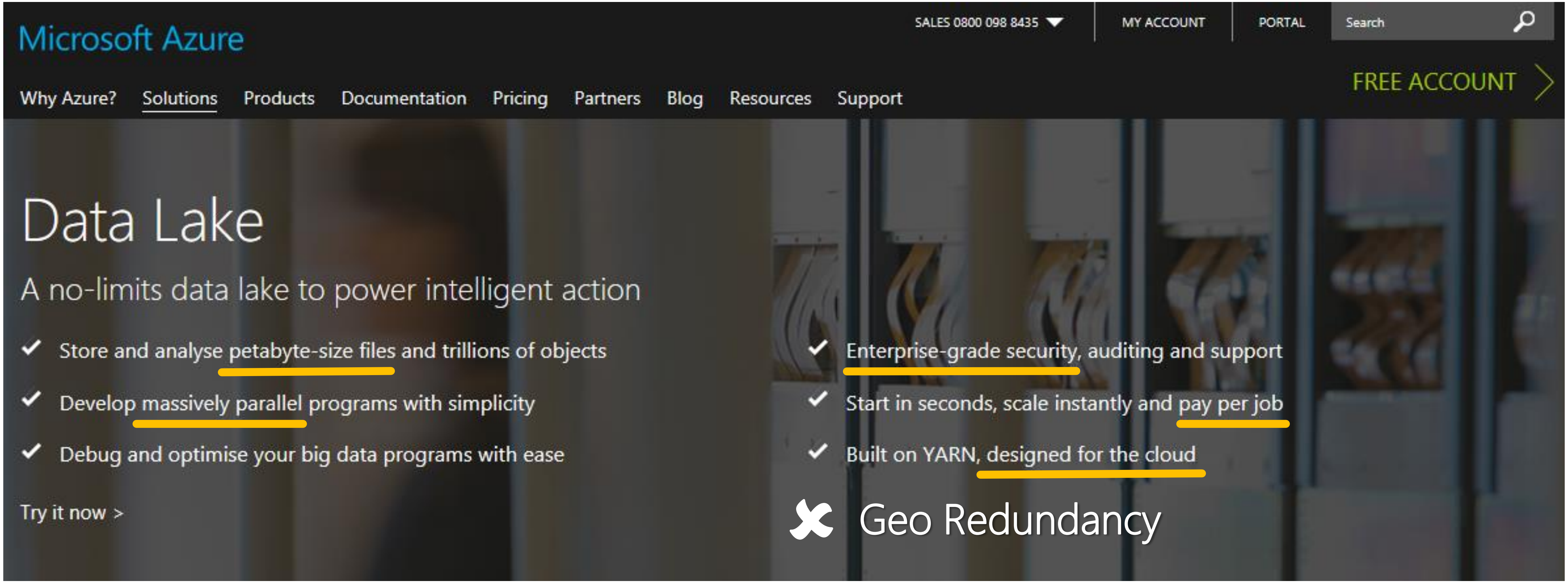
Development & Management

U-SQL

‘Hello World’ to Advanced Analytics

Why use Azure Data Lake?

The Microsoft version:



Data Lake

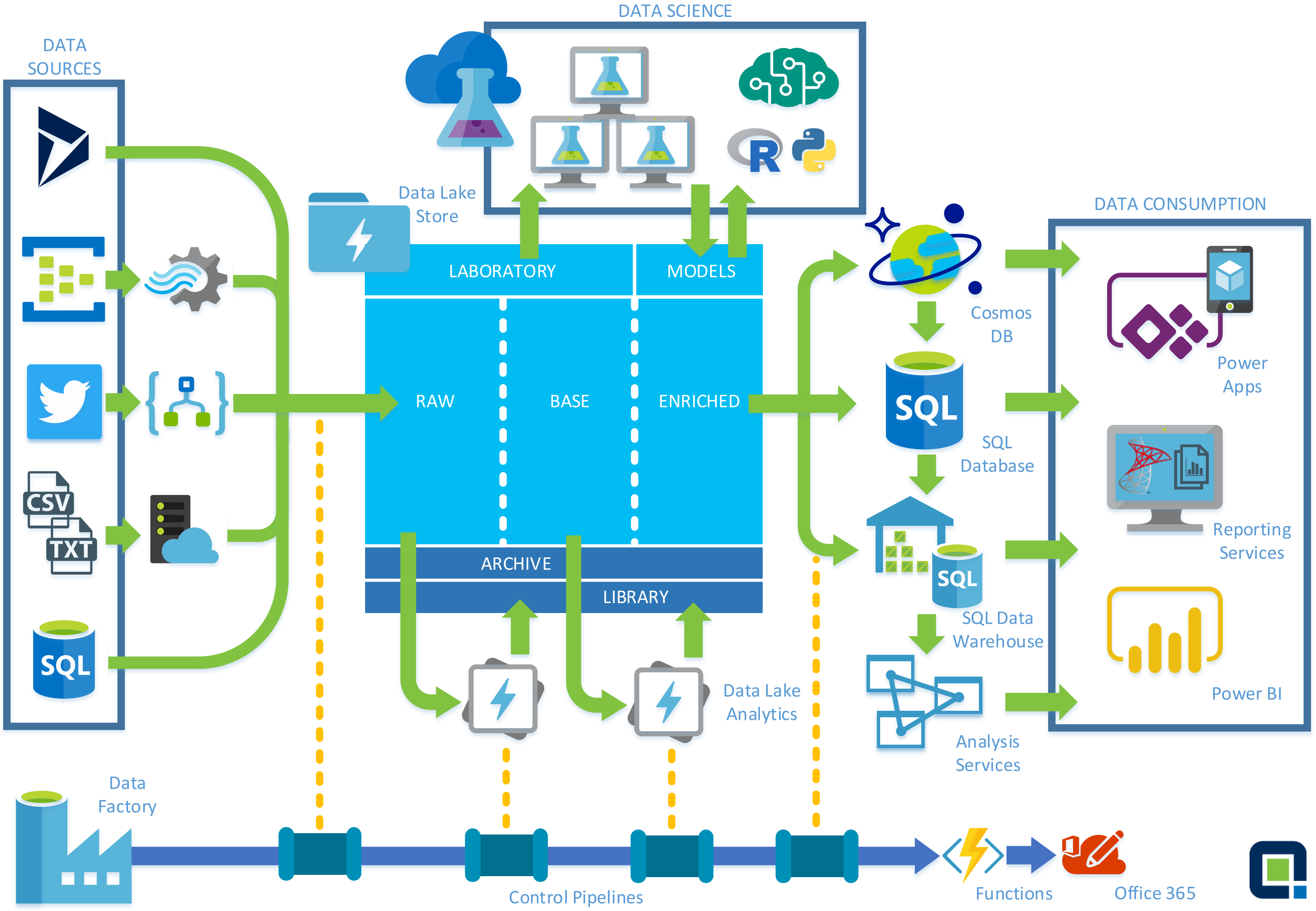
A no-limits data lake to power intelligent action

- ✓ Store and analyse petabyte-size files and trillions of objects
- ✓ Develop massively parallel programs with simplicity
- ✓ Debug and optimise your big data programs with ease
- ✓ Enterprise-grade security, auditing and support
- ✓ Start in seconds, scale instantly and pay per job
- ✓ Built on YARN, designed for the cloud
- ✗ Geo Redundancy

Try it now >

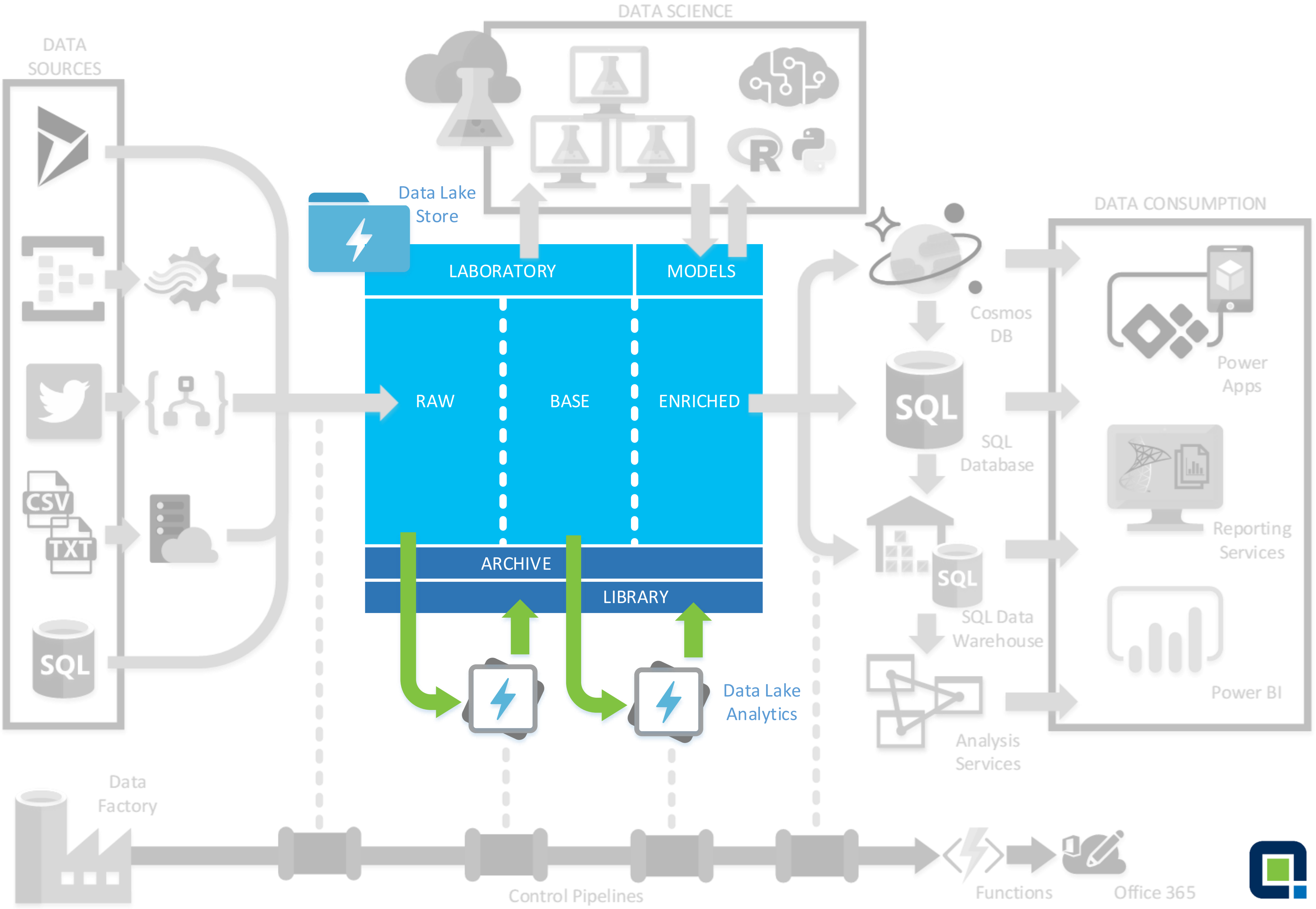
https://azure.microsoft.com/en-gb/solutions/data-lake/

The Modern Data Warehouse



adatis

The Modern Data Warehouse



adatis

Agenda

What is Azure Data Lake?

Storage & Compute

Why use Data Lake?

The Modern Data Warehouse

How can we work with Data Lake?

Development & Management

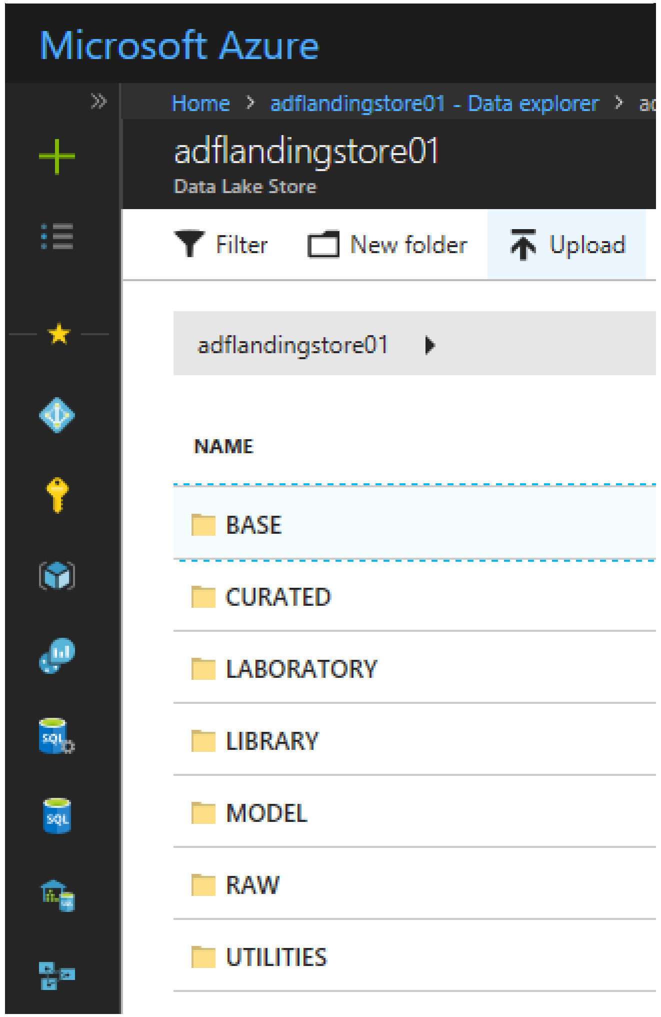
U-SQL

‘Hello World’ to Advanced Analytics

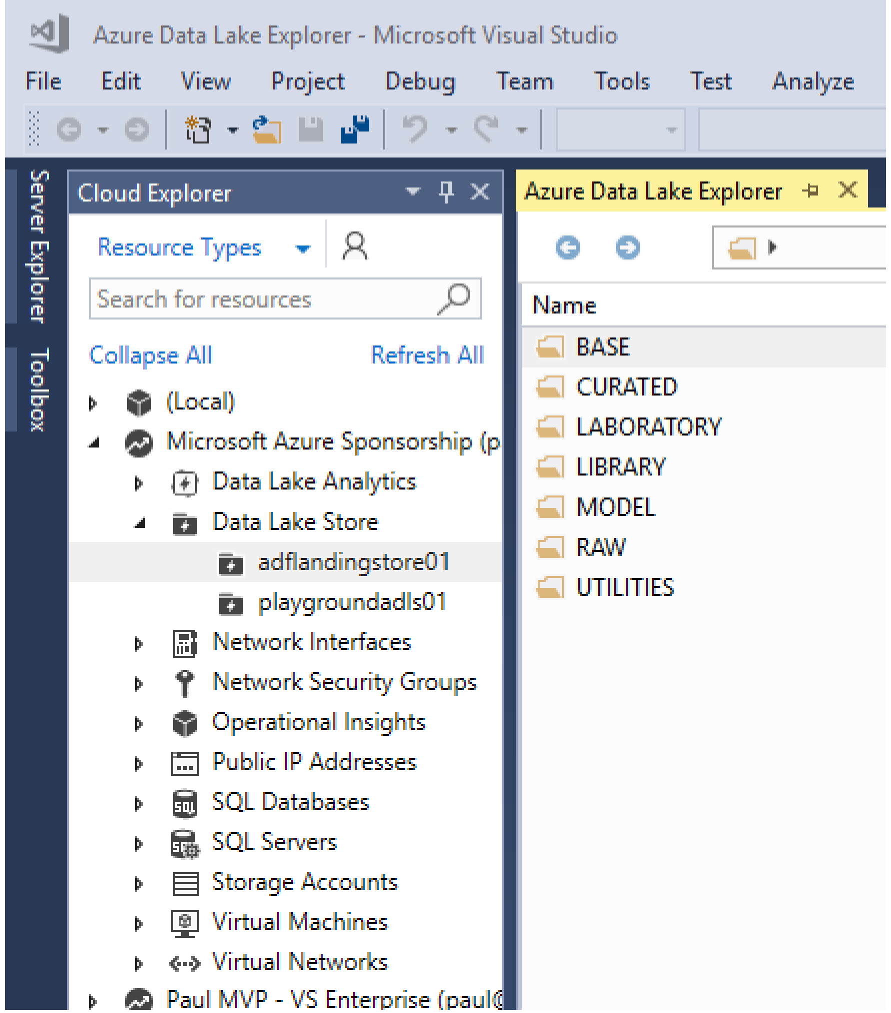
Working with Azure Data Lake Storage

Manual File Uploads

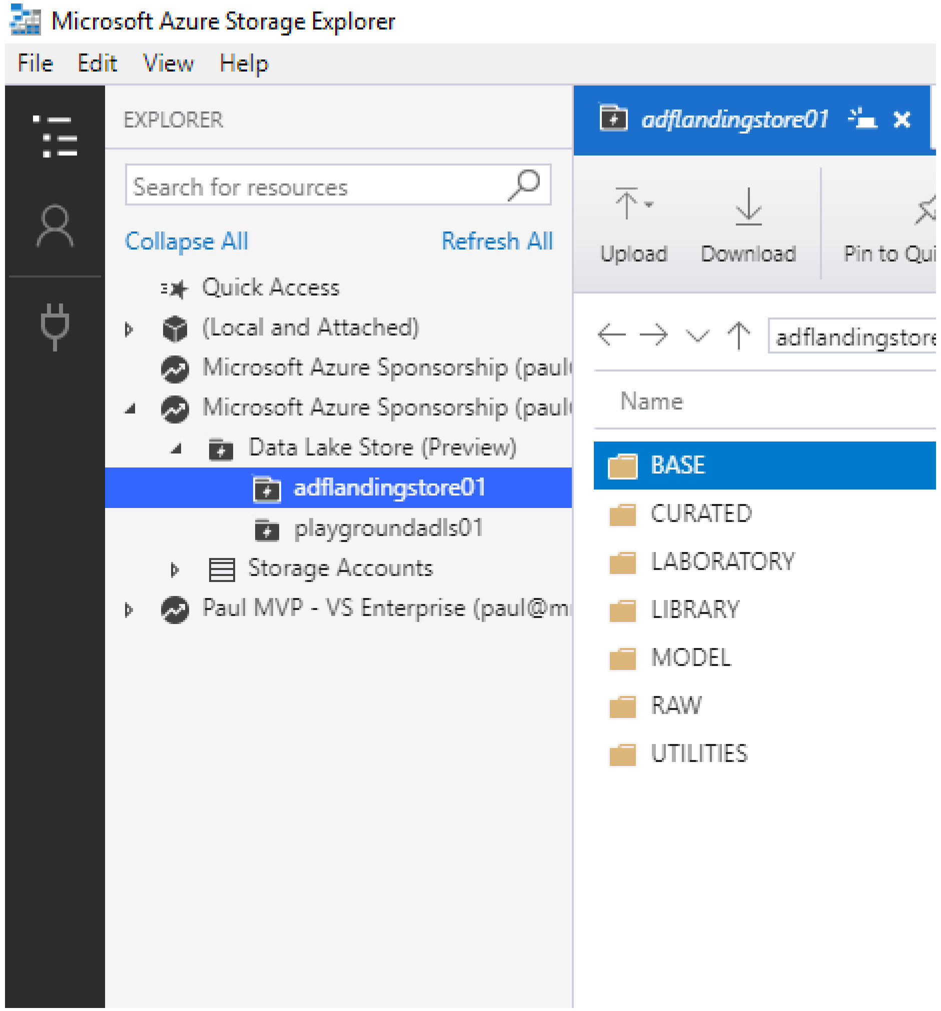
Azure Portal



Visual Studio Cloud Explorer



Azure Storage Explorer



Working with Azure Data Lake Storage

Automatic File Uploads

Data Factory

Stream Analytics

Logic Apps

SSIS

.Net SDK

PowerShell

Python

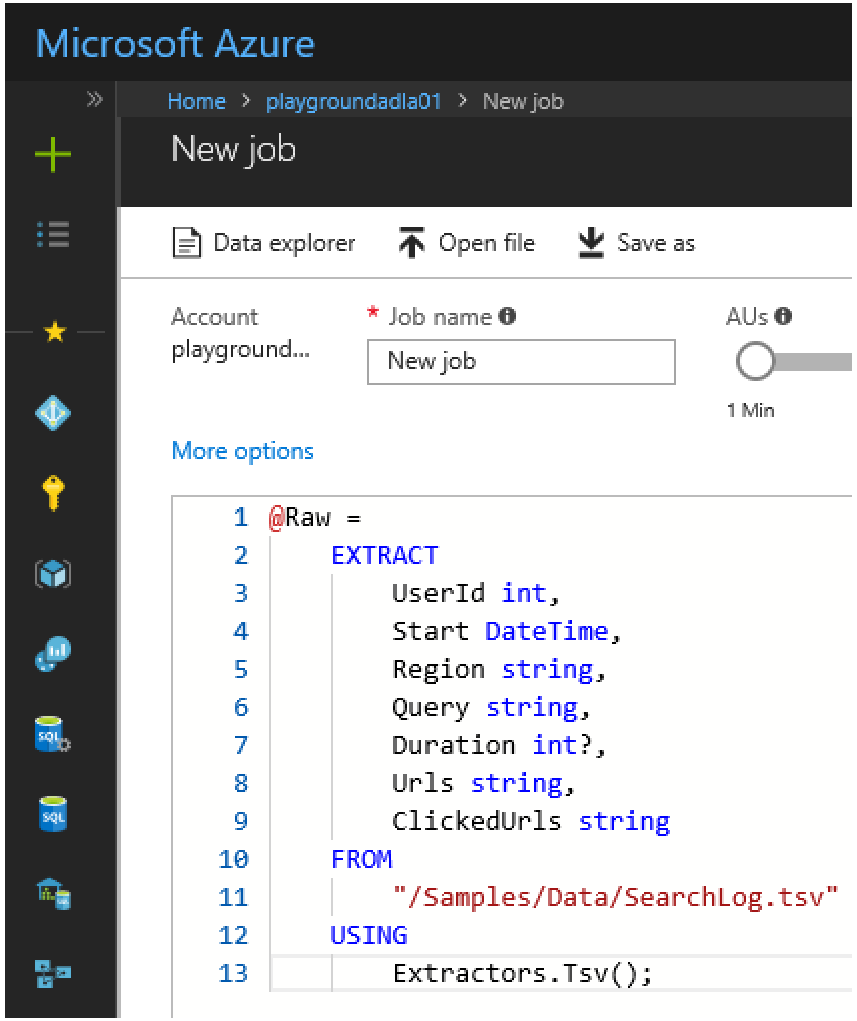
REST API

✗ VNet

Working with Azure Data Lake Analytics

Manual U-SQL Job Execution

Azure Portal



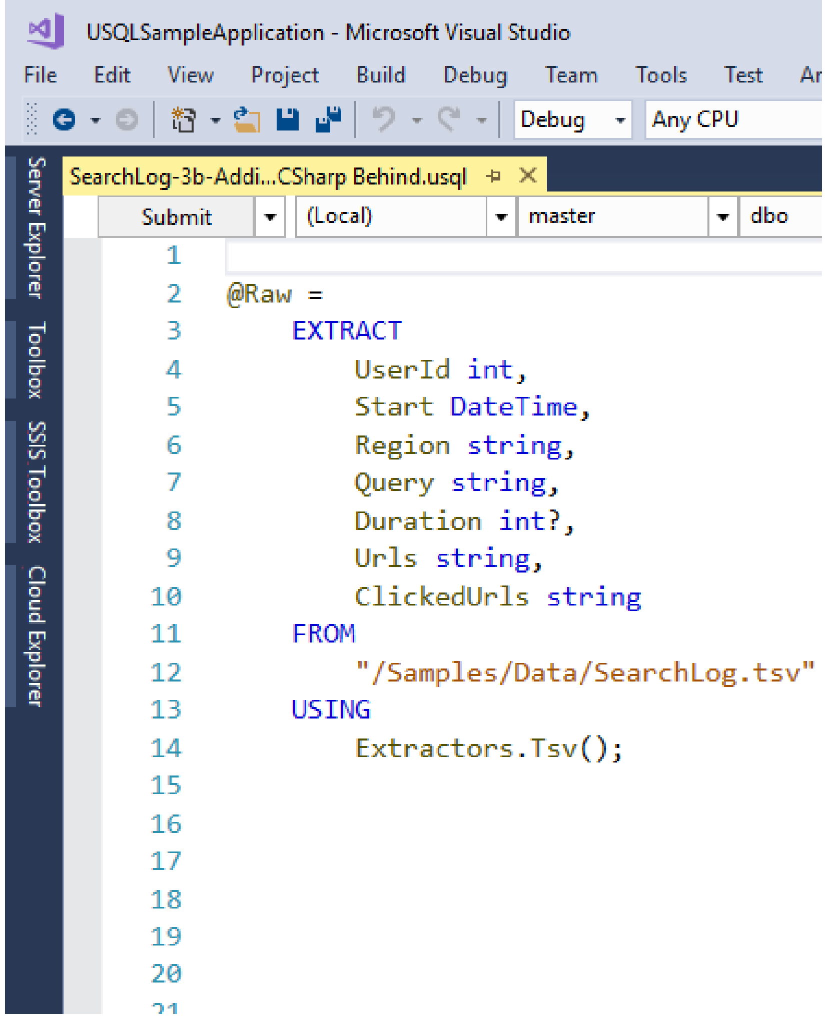
Account: playground...

* Job name: New job

AUs: 1 Min

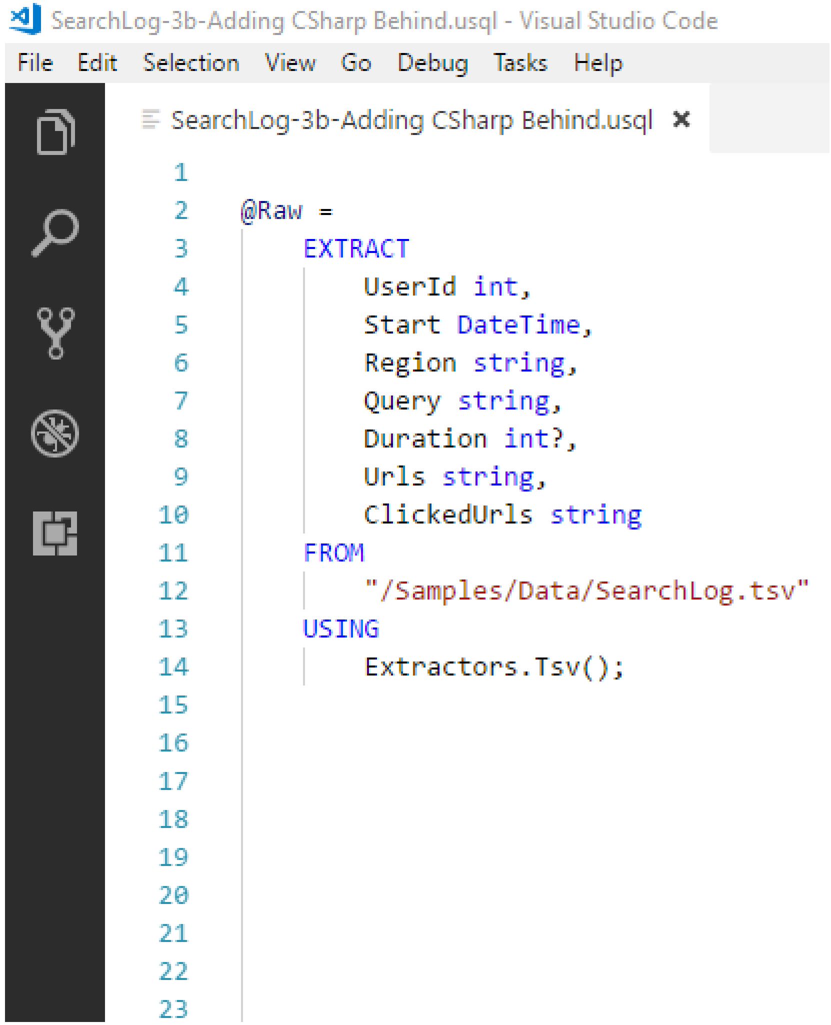
```
@Raw =
    EXTRACT
        UserId int,
        Start DateTime,
        Region string,
        Query string,
        Duration int?,
        Urls string,
        ClickedUrls string
    FROM
        "/Samples/Data/SearchLog.tsv"
    USING
        Extractors.Tsv();
```

Visual Studio Project



```

@Raw =
    EXTRACT
        UserId int,
        Start DateTime,
        Region string,
        Query string,
        Duration int?,
        Urls string,
        ClickedUrls string
    FROM
        "/Samples/Data/SearchLog.tsv"
    USING
        Extractors.Tsv();
```

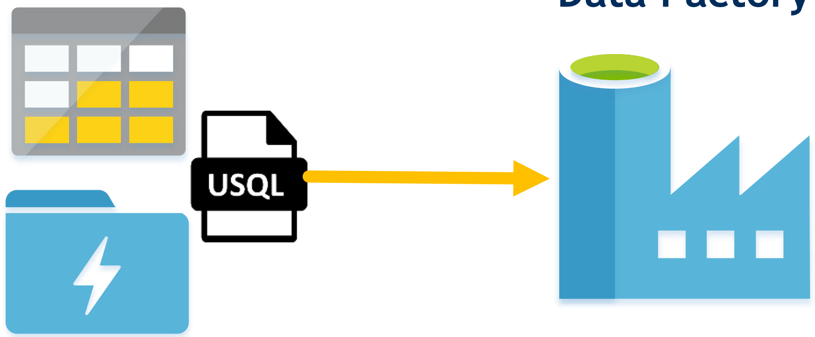
Visual Studio Code


SearchLog-3b-Adding CSharp Behind.usql - Visual Studio Code
File
Edit
Selection
View
Go
Debug
Tasks
Help
SearchLog-3b-Adding CSharp Behind.usql

```

@Raw =
    EXTRACT
        UserId int,
        Start DateTime,
        Region string,
        Query string,
        Duration int?,
        Urls string,
        ClickedUrls string
    FROM
        "/Samples/Data/SearchLog.tsv"
    USING
        Extractors.Tsv();
```

Working with Azure Data Lake Analytics

Automatic U-SQL Job Execution



Data Factory

.Net SDK

PowerShell

Azure CLI

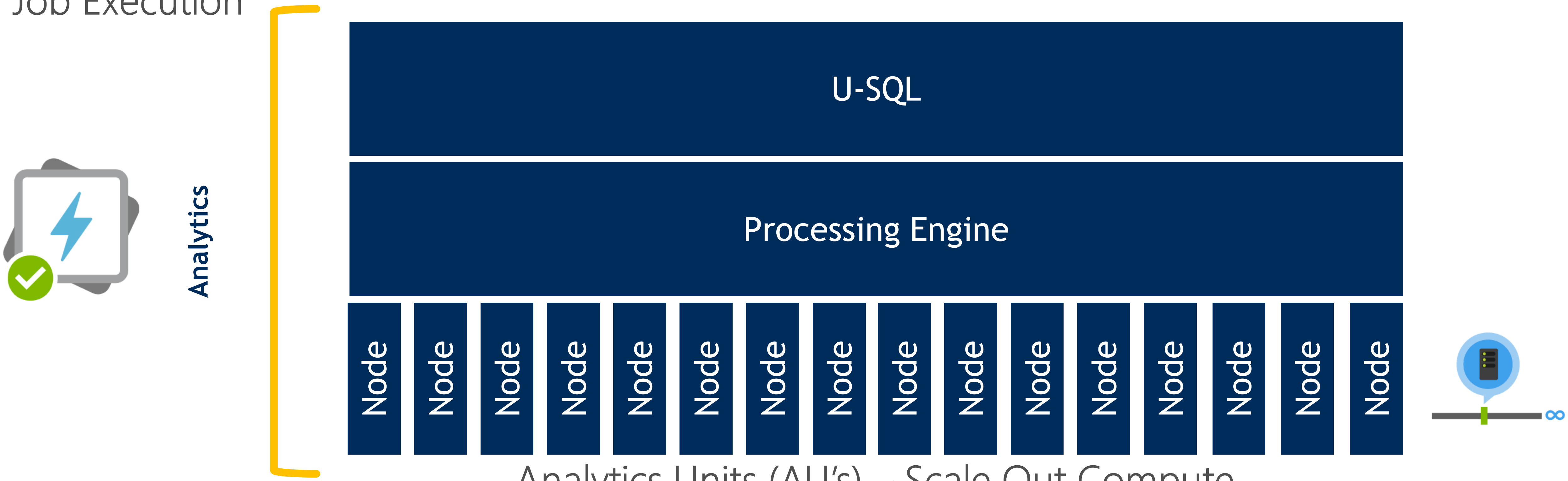
Python

Java

Node.js

Working with Azure Data Lake Analytics

Job Execution



Analytics

U-SQL

Processing Engine

Node Node Node Node Node Node Node Node Node Node Node Node Node Node Node Node

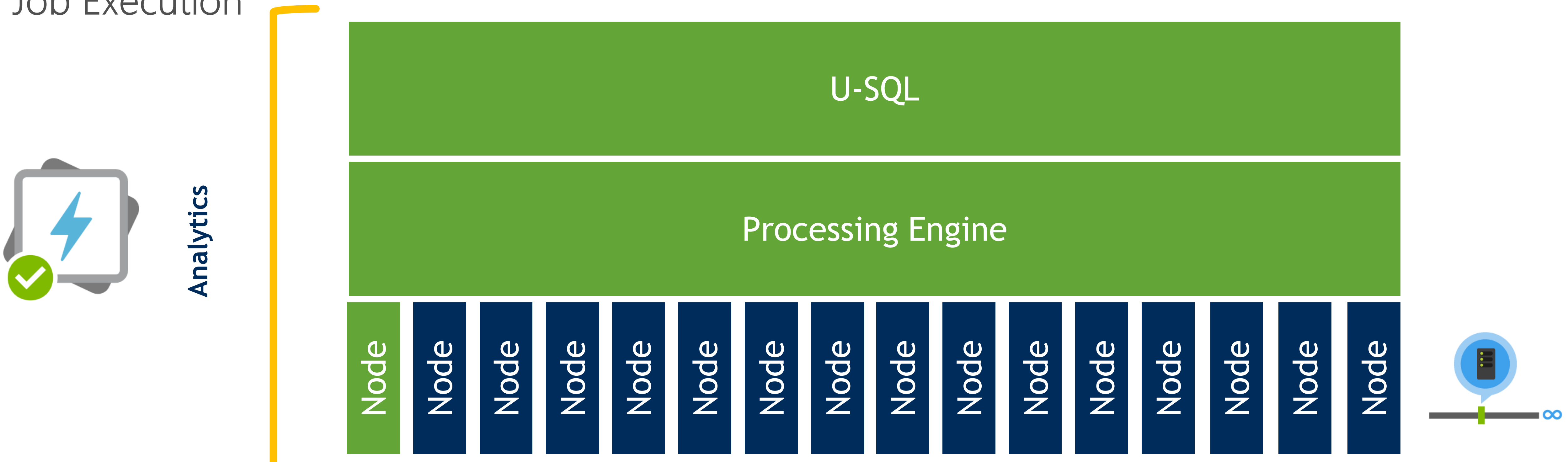
∞

Analytics Units (AU’s) – Scale Out Compute

AU/hours

Working with Azure Data Lake Analytics

Job Execution



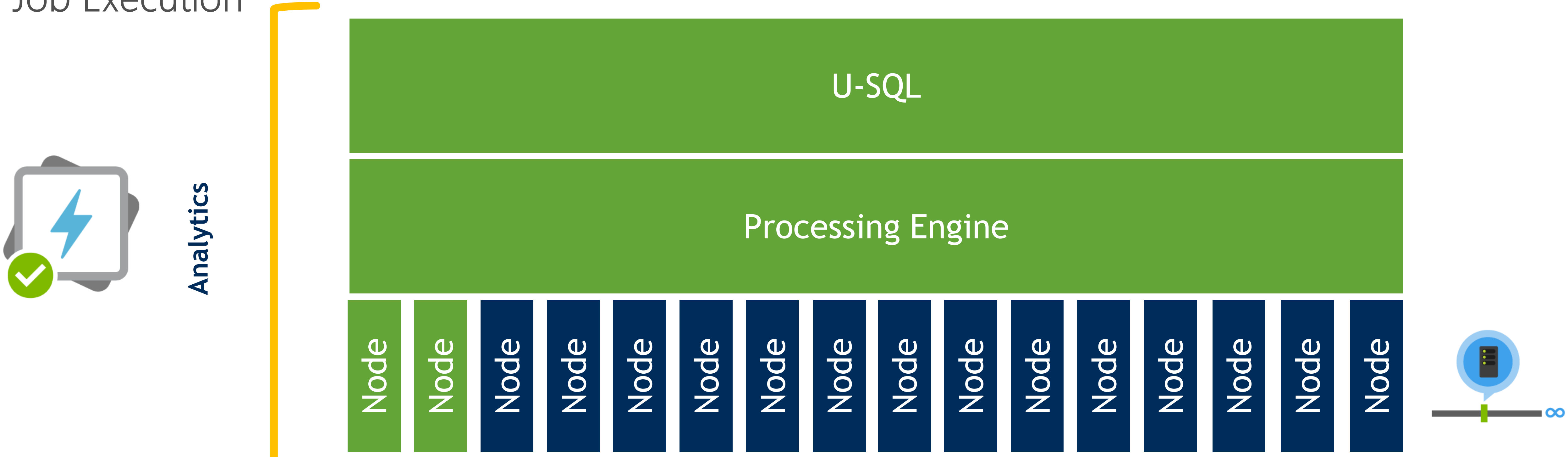
Analytics Units (AU’s) – Scale Out Compute

1 x AU/hour = £1.49 *

* Price Checked April 2018

Working with Azure Data Lake Analytics

Job Execution



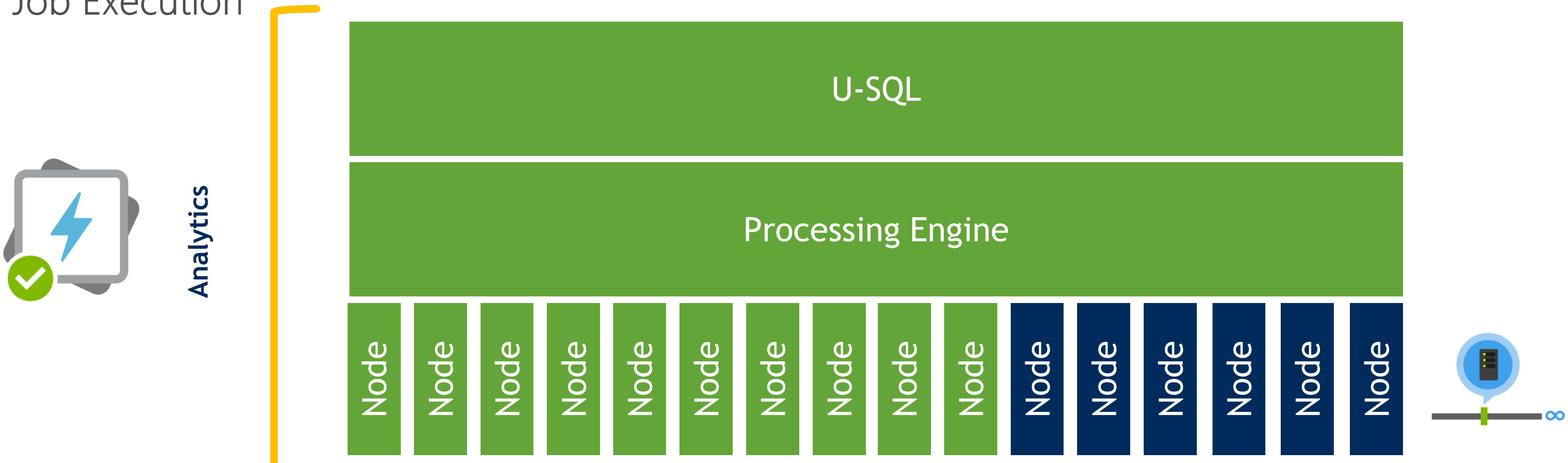
Analytics Units (AU's) – Scale Out Compute

2 x AU/hour = £2.98 *

* Price Checked April 2018

Working with Azure Data Lake Analytics

Job Execution



Analytics

U-SQL

Processing Engine

Node Node Node Node Node Node Node Node Node Node Node Node Node Node Node Node

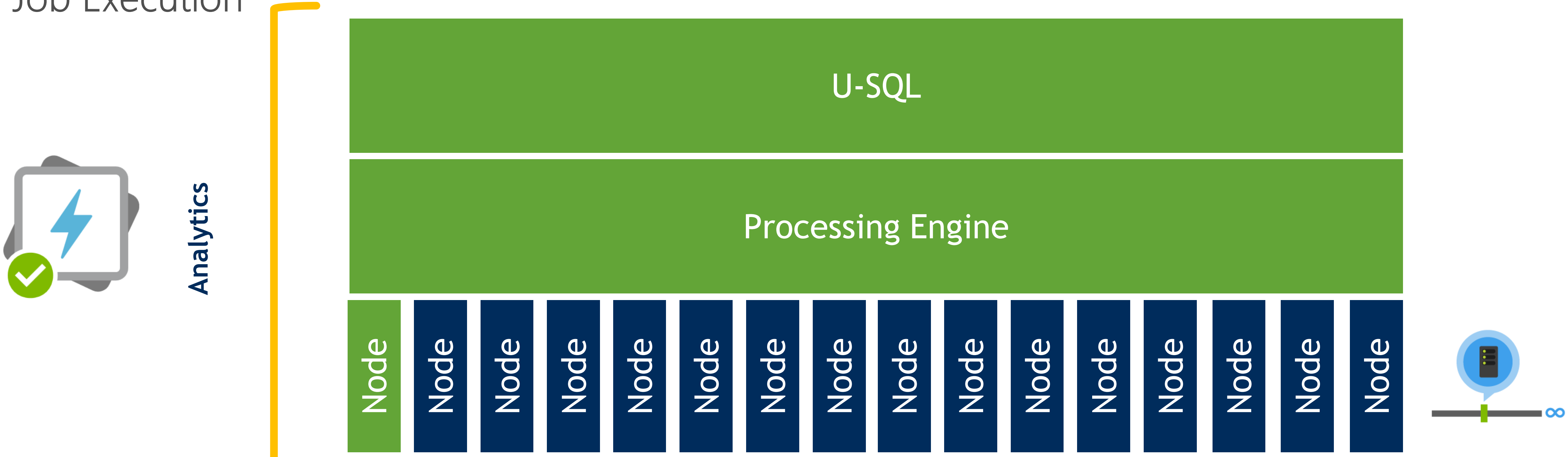
Analytics Units (AU's) – Scale Out Compute

10 x AU/hour = £14.90 *

* Price Checked April 2018

Working with Azure Data Lake Analytics

Job Execution



Analytics

U-SQL

Processing Engine

Node Node Node Node Node Node Node Node Node Node Node Node Node Node Node Node

Analytics Units (AU's) – Scale Out Compute

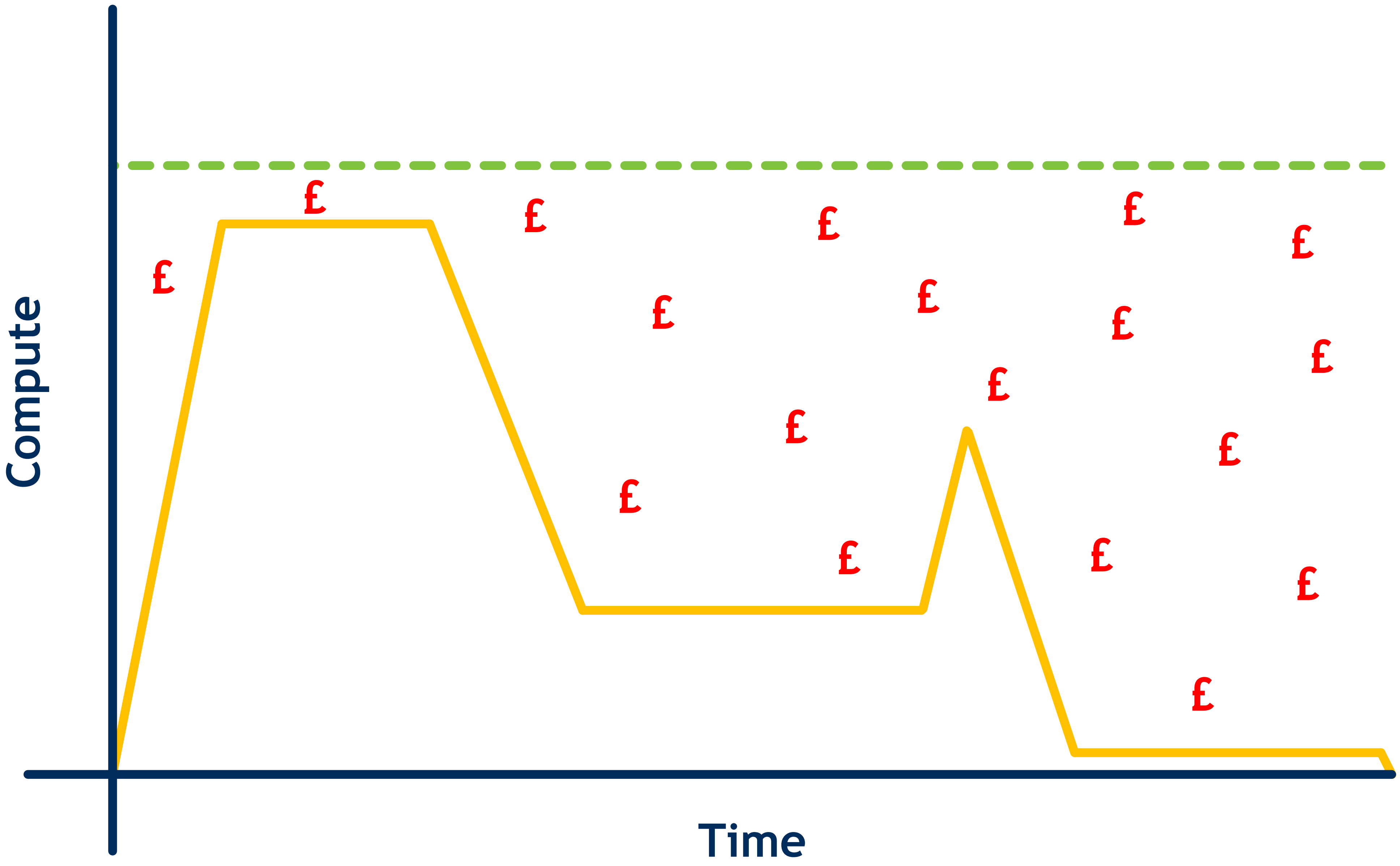
1 minute to complete

1 x AU/hour = ~~£1.49~~ *

£0.02

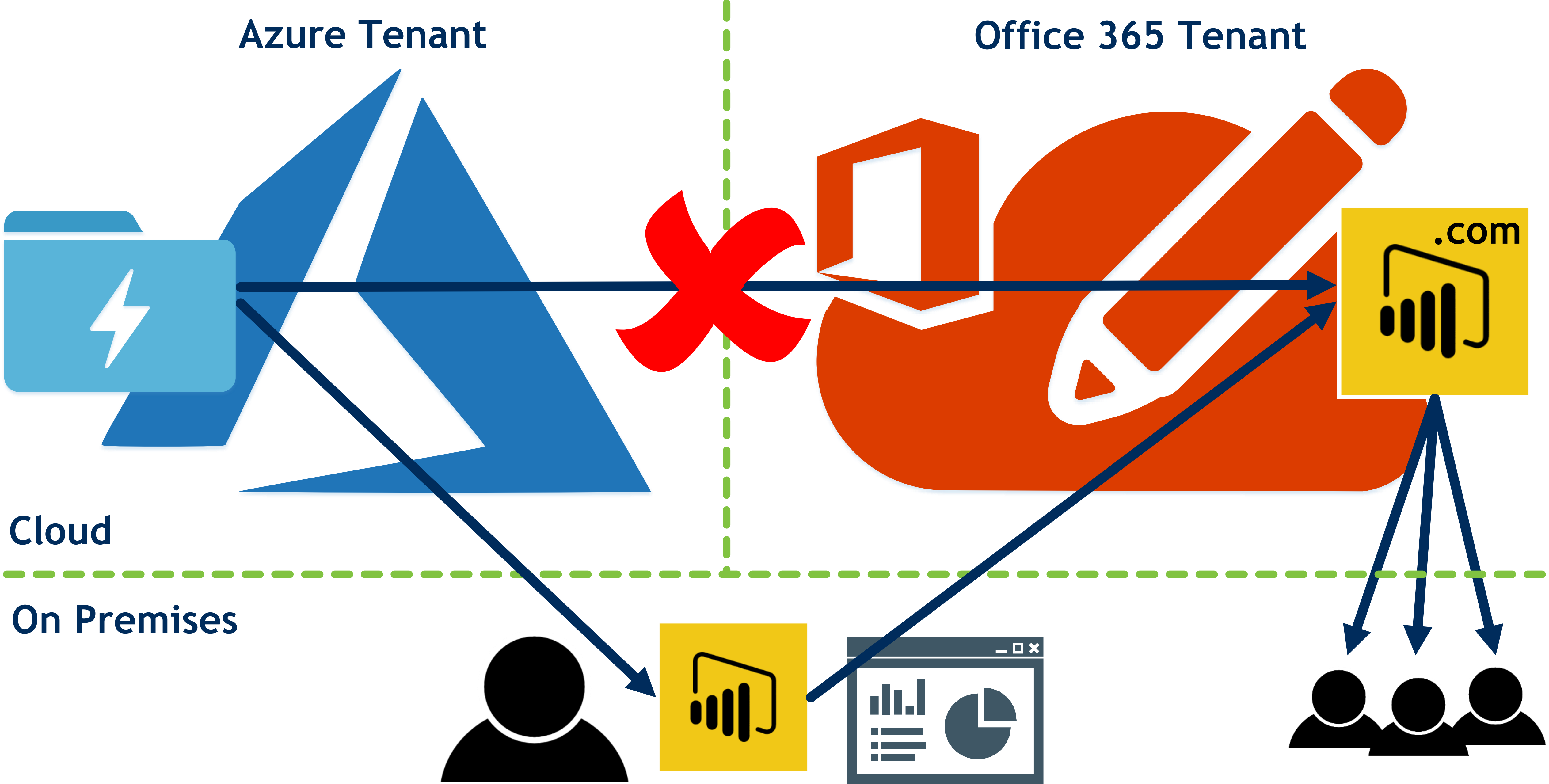
* Price Checked April 2018

Working with Azure Data Lake Analytics



adatis

Consuming Azure Data Lake



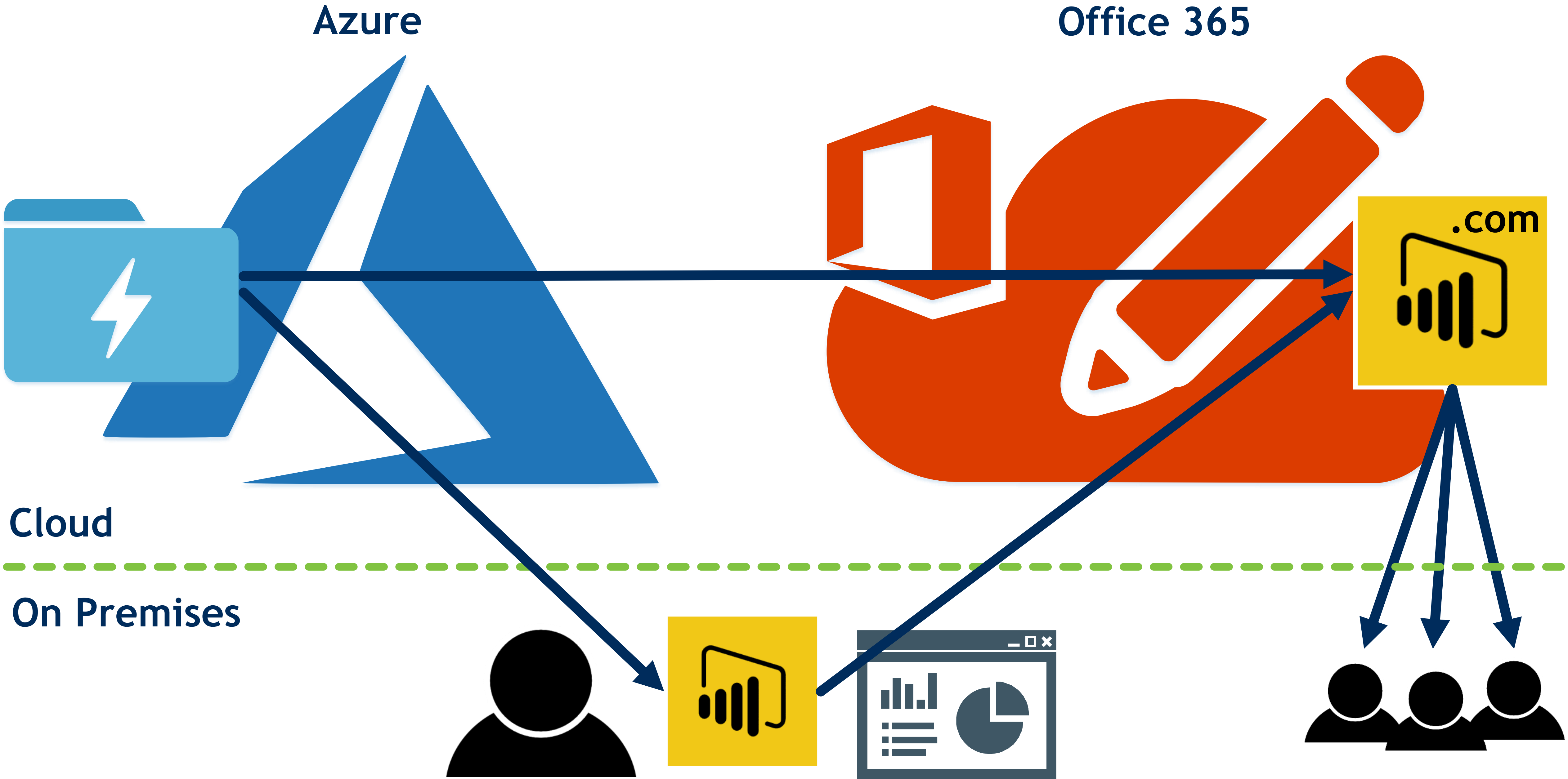
Azure Tenant

Office 365 Tenant

Cloud

On Premises

Consuming Azure Data Lake



Agenda

What is Azure Data Lake?

Storage & Compute

Why use Data Lake?

The Modern Data Warehouse

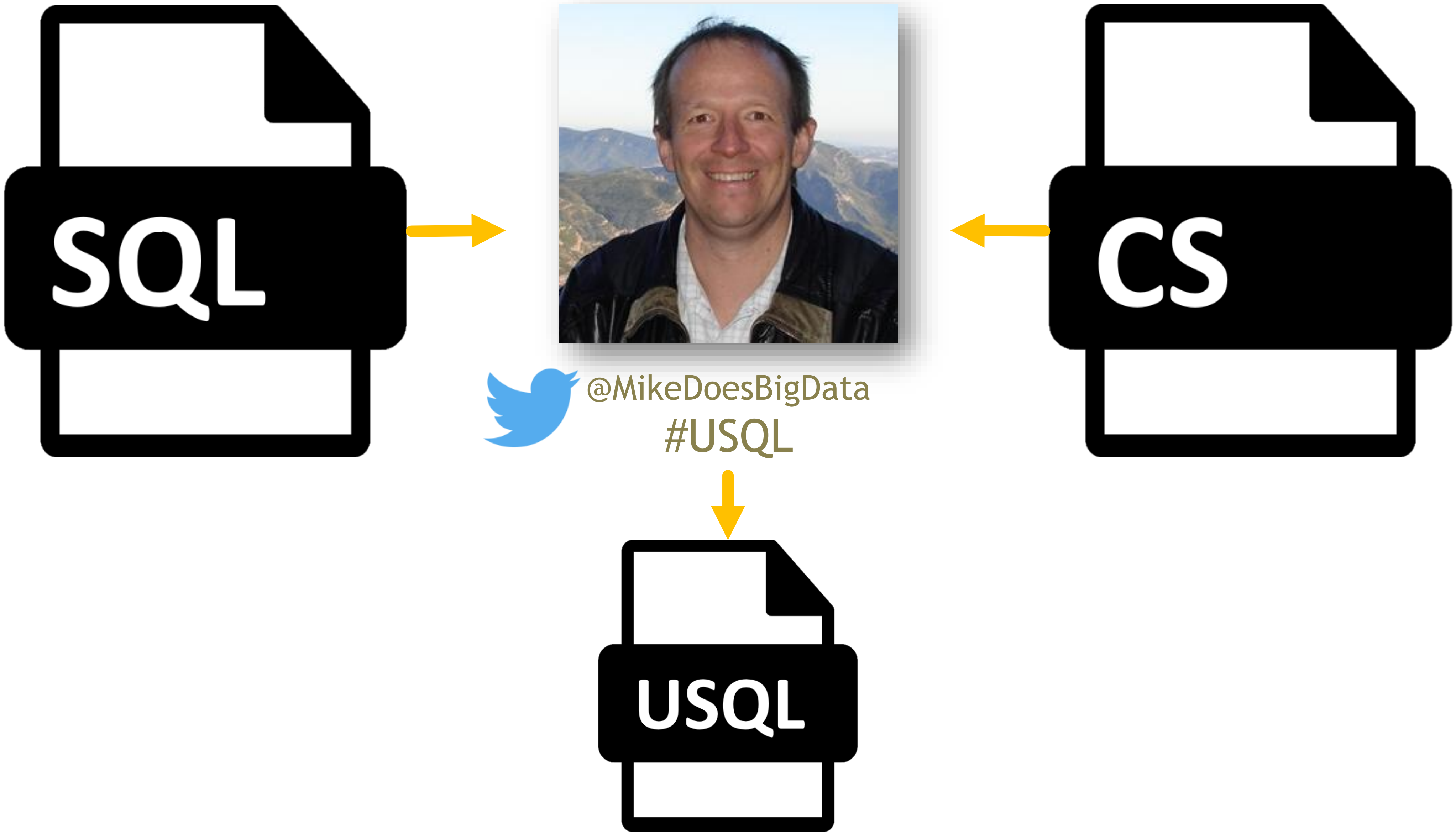
How can we work with Data Lake?

Development & Management

U-SQL

‘Hello World’ to Advanced Analytics

What is U-SQL?

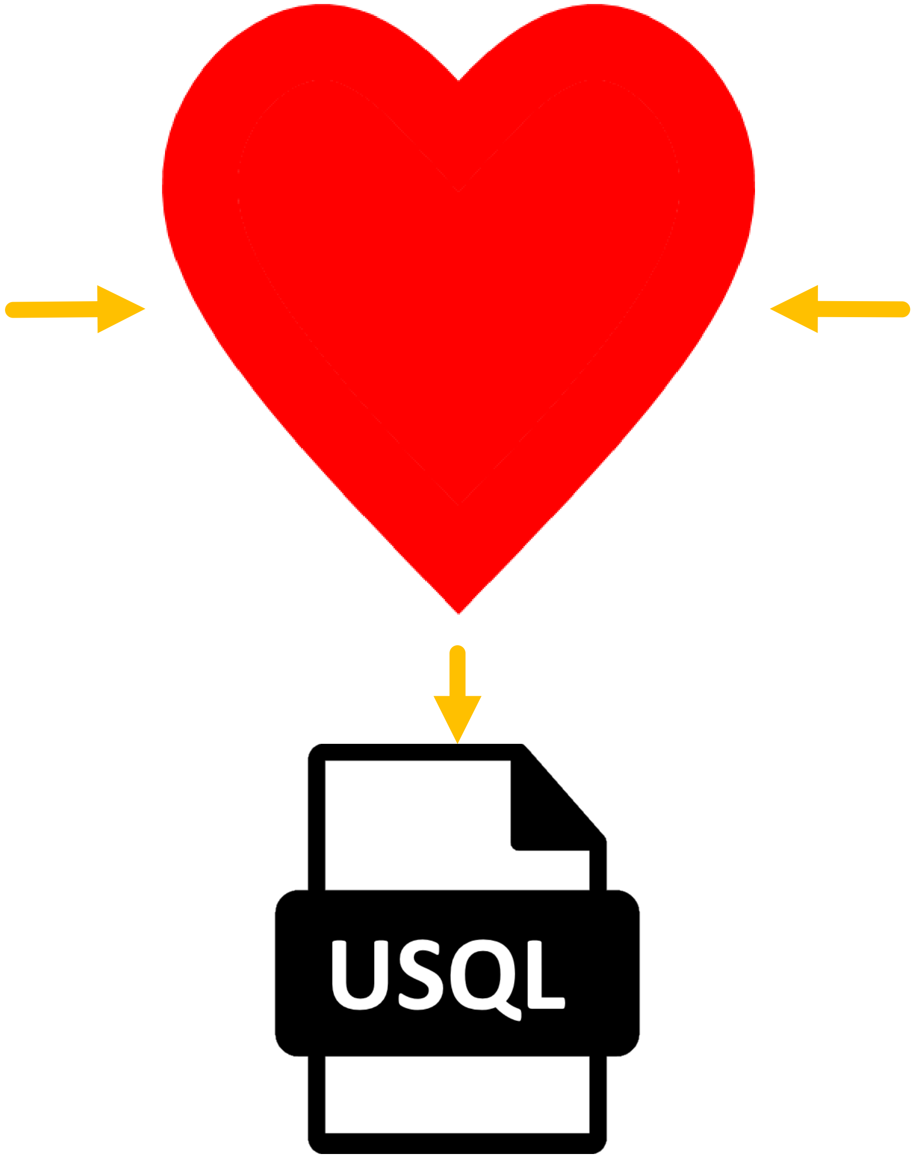


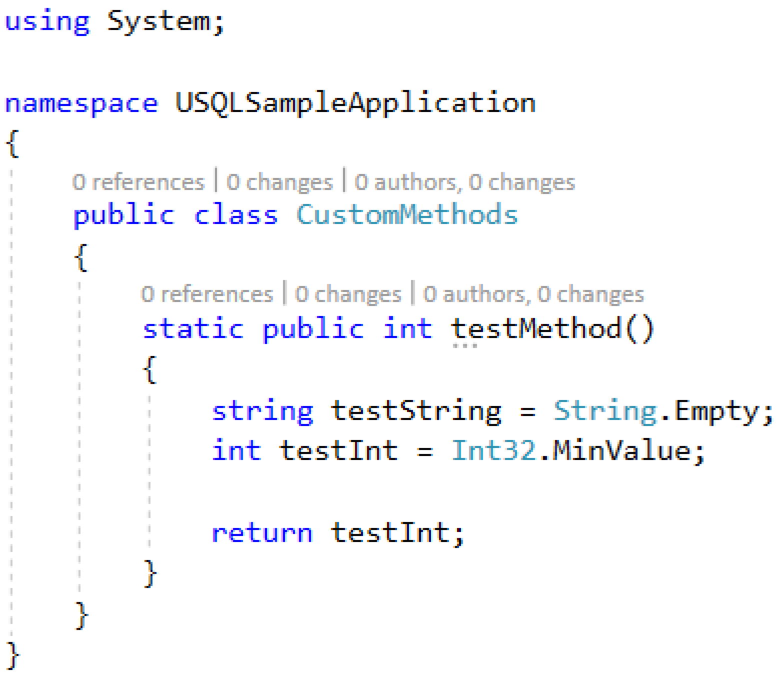
@MikeDoesBigData

#USQL

What is U-SQL?

```
SELECT
     Domain,
     COUNT(*) AS Qty
FROM
     @Domains
GROUP BY
     Domain;
```

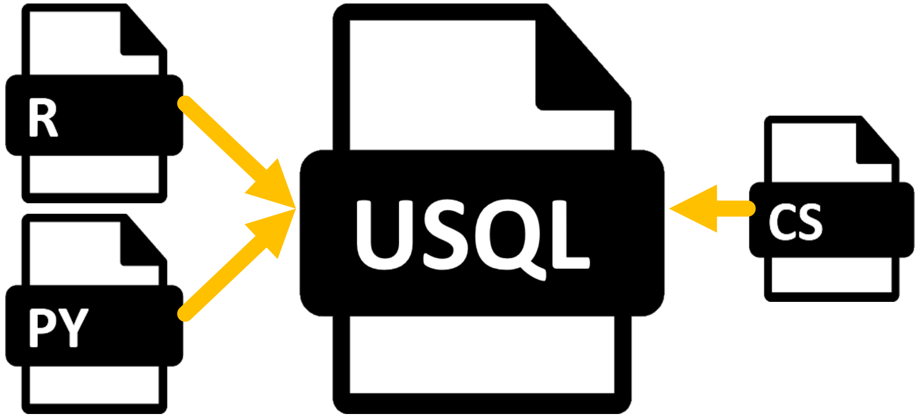




```
using System;

namespace USQLSampleApplication
{
    0 references | 0 changes | 0 authors, 0 changes
    public class CustomMethods
    {
        0 references | 0 changes | 0 authors, 0 changes
        static public int testMethod()
        {
            string testString = String.Empty;
            int testInt = Int32.MinValue;

            return testInt;
        }
    }
}
```

What is U-SQL?

```
@SizeAndCount =
    SELECT
        [ModifiedDate].ToString("yyyy") AS Year,
        [FileName].Substring([FileName].IndexOf(".") + 1, 3) AS FileExtension,
        COUNT(0) AS RecordCount,
        Math.Ceiling(Convert.ToDecimal(SUM([Size]))) AS FileSizeTotalsMB,
        Math.Ceiling(Convert.ToDecimal(SUM([Size])/1024)) AS FileSizeTotalsGB
    FROM
        @Raw
    WHERE
        [ActualFileName] == "FileDetailsTest.csv"
    GROUP BY
        [ModifiedDate].ToString("yyyy"),
        [FileName].Substring([FileName].IndexOf(".") + 1, 3);
```



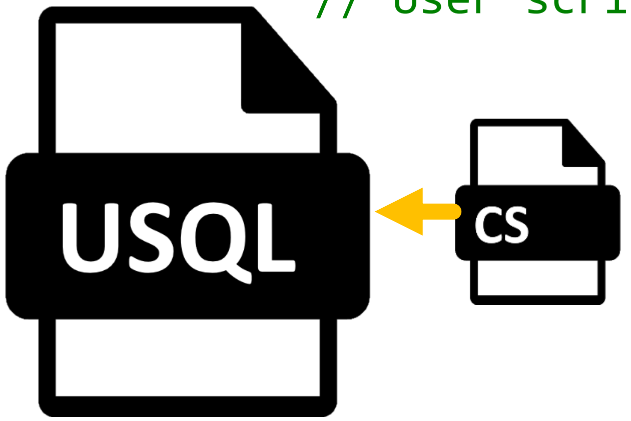
U-SQL and C# Code Behind

Automatic – Stored Procedures

```
// Assemblies from class library

CREATE ASSEMBLY IF NOT EXISTS [Mine]
FROM @"\CustomStringMethods.dll";


// User code wrapped in proc

CREATE PROCEDURE StoredProc01()
AS
BEGIN

    REFERENCE ASSEMBLY [Mine];

END;
```

Manual – Adhoc Job Submit

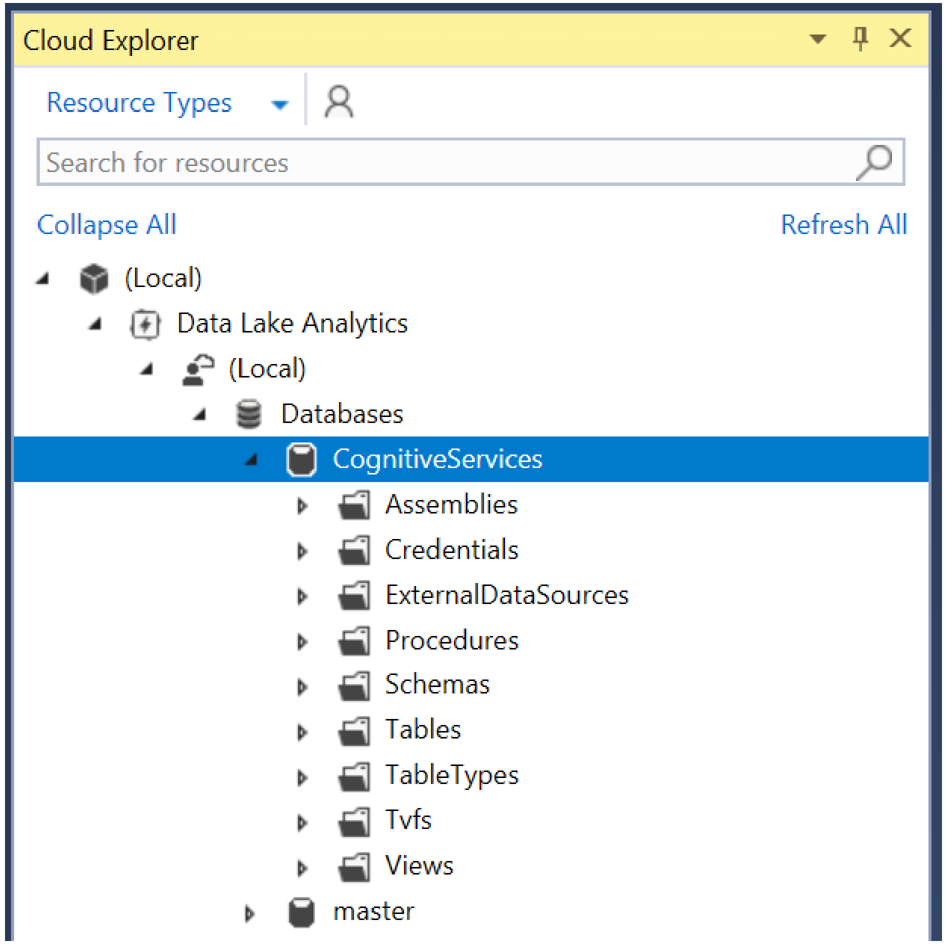
```
// Auto-generated header code
// Generated Code Behind Header

CREATE ASSEMBLY [__codeBehind_1xkprrnp.trv]
FROM 0x4D5A9000030000000400000000;

REFERENCE ASSEMBLY [__codeBehind_1xkprrnp.trv];


// Generated Code Behind Header
// Auto-generated header code ended
// User script
```



```
DROP ASSEMBLY
[__codeBehind_1xkprrnp.trv];
```

Data Lake Analytics Database

catalog



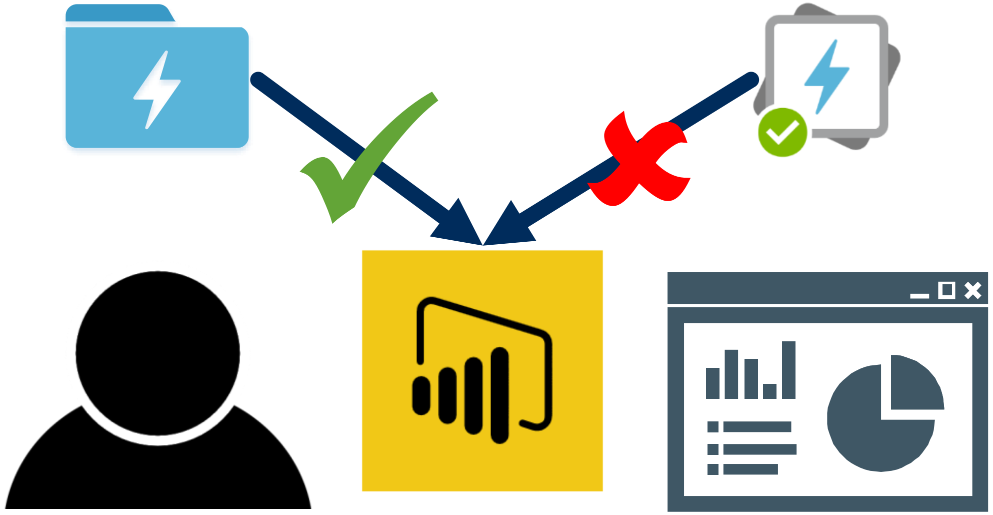
Cloud Explorer

Resource Types

Search for resources

Collapse All　　Refresh All

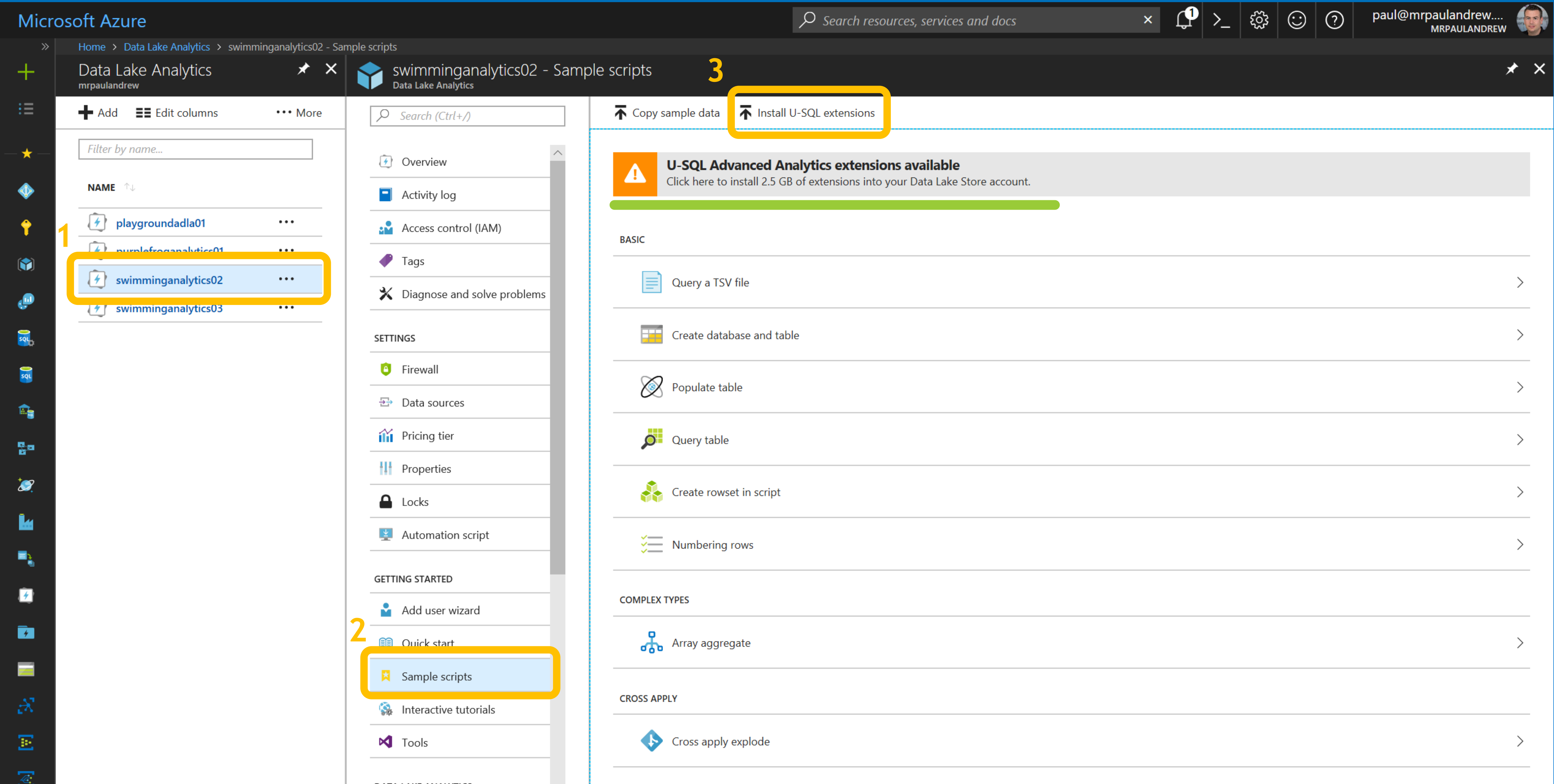
- (Local)
  - Data Lake Analytics
    - (Local)
      - Databases
        - CognitiveServices
          - Assemblies
          - Credentials
          - ExternalDataSources
          - Procedures
          - Schemas
          - Tables
          - TableTypes
          - Tvfs
          - Views
        - master



```
CREATE DATABASE IF NOT EXISTS BaseOfData01
CREATE ASSEMBLY IF NOT EXISTS ImageCommon
CREATE PROCEDURE IF NOT EXISTS StoredProc01
CREATE SCHEMA IF NOT EXISTS Schema01
CREATE TABLE IF NOT EXISTS Table01
CREATE VIEW IF NOT EXISTS View01
```

adatis

Getting the U-SQL Extensions

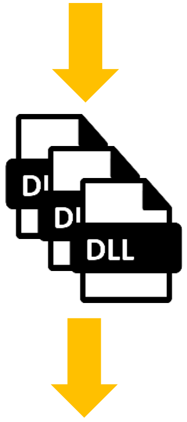


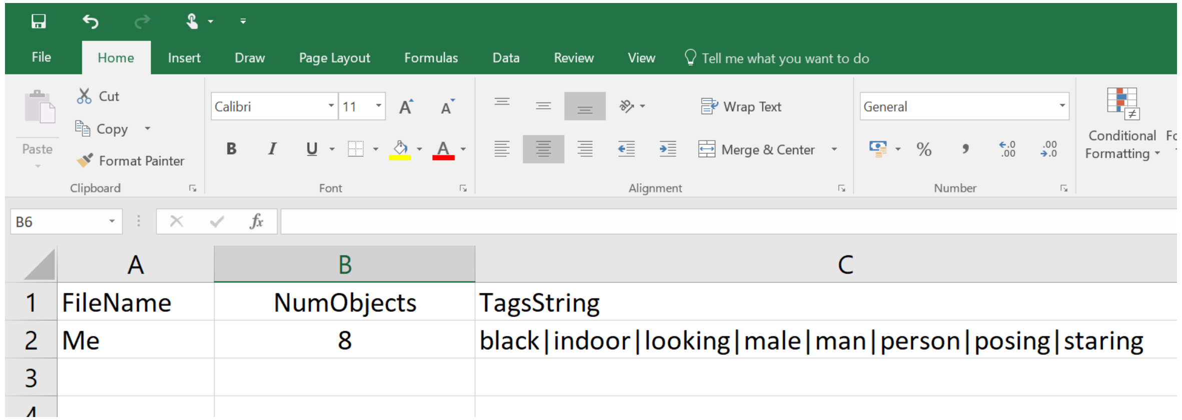
U-SQL Advanced Analytics extensions available

Click here to install 2.5 GB of extensions into your Data Lake Store account.

U-SQL Image Tagging with Cognitive Services

```
// Load Assemblies
REFERENCE ASSEMBLY ImageCommon;
REFERENCE ASSEMBLY FaceSdk;
REFERENCE ASSEMBLY ImageEmotion;
REFERENCE ASSEMBLY ImageTagging;
REFERENCE ASSEMBLY ImageOcr;

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

//Tagging processor
@tags_from_processor =
    PROCESS @imgs
    PRODUCE FileName, NumObjects int, Tags SQL.MAP<string, float?>
    READONLY FileName USING new Cognition.Vision.ImageTagger();

@tags_from_processor_serialized =
    SELECT
        FileName,
        NumObjects,
        String.Join
        ("|", Tags.Select(x => String.Format("{0}", x.Key))) AS TagsString
    FROM
        @tags_from_processor;

//Output
OUTPUT @tags_from_processor_serialized
TO @"/Output/FileTags.csv"
USING Outputters.Csv(outputHeader : true);
```





	A	B	C
1	FileName	NumObjects	TagsString
2	Me	8	black|indoor|looking|male|man|person|posing|staring
3			
4			

Further Reading

Microsoft U-SQL Language Reference Guide
https://msdn.microsoft.com/en-us/azure/data-lake-analytics/u-sql/u-sql-language-reference

SQL Server Central Stairway (21 chapters)
http://www.sqlservercentral.com/stairway/142480/

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

MrPaulAndrew.com
https://mrpaulandrew.com/tag/u-sql/

Adatis Blogs
http://blogs.adatis.co.uk/search?q=U-SQL

Thanks for Listening

Paul Andrew

@MrPaulAndrew

MVP Microsoft® Most Valuable Professional

adatis

Blog: http://mrpaulandrew.com
Email: paul@mrpaulandrew.com