SeaScale Meetup
Jan 2016

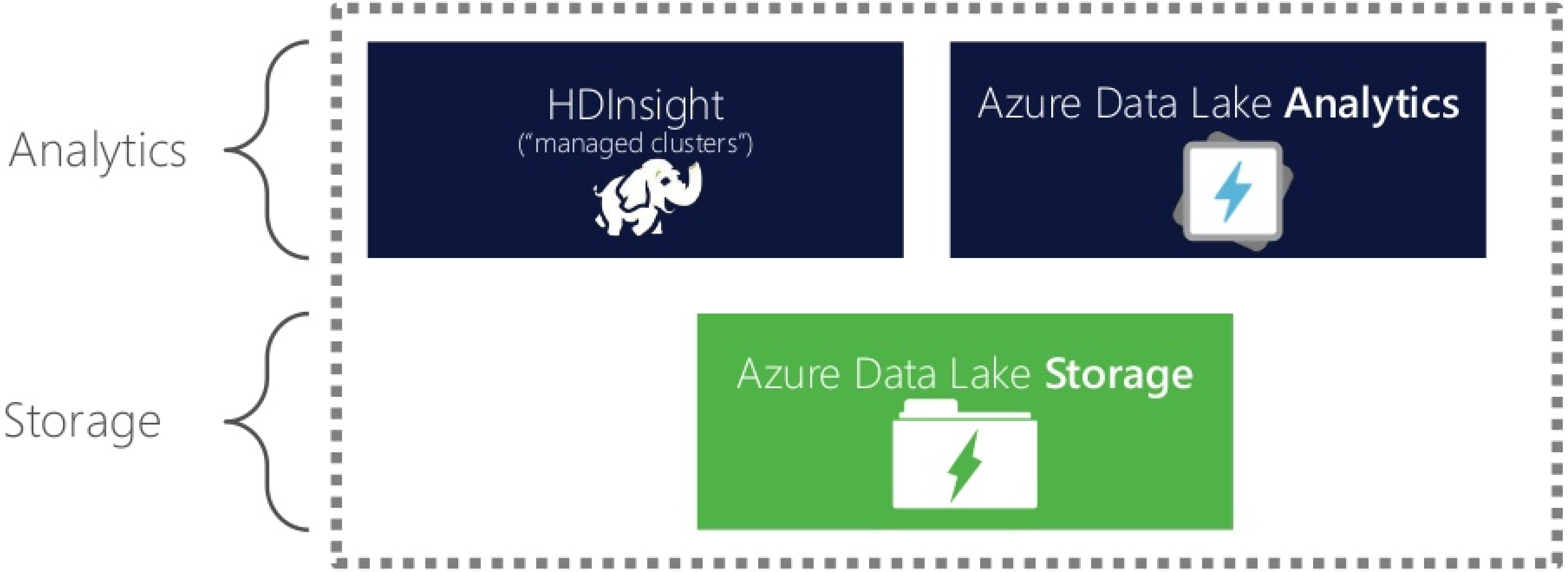
Azure Data Lake & U-SQL

Michael Rys, @MikeDoesBigData

http://www.azure.com/datalake
{mrys, usql}@microsoft.com

Microsoft

Azure Data Lake



Analytics

HDInsight
("managed clusters")

Azure Data Lake **Analytics**

Storage

Azure Data Lake **Storage**

ADLA complements HDInsight

Target the same scenarios, tools, and customers

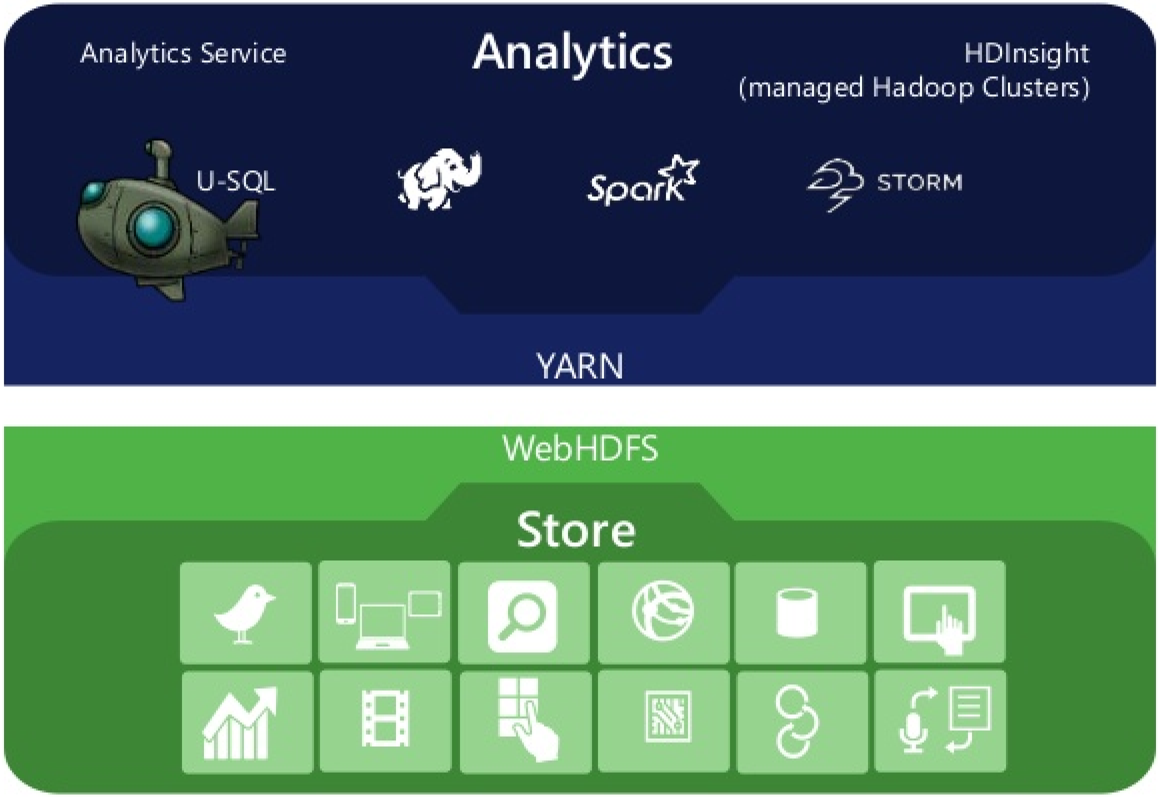
HDInsight

- For developers familiar with the Open Source: Java, Eclipse, Hive, etc.
- Clusters offer customization, control, and flexibility in a managed Hadoop cluster

ADLA

- Enables customers to leverage existing experience with C#, SQL & PowerShell
- Offers convenience, efficiency, automatic scale, and management in a "job service" form factor

Azure Data Lake



Microsoft

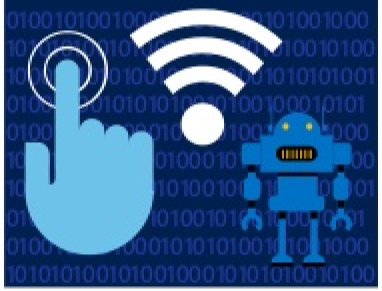
Azure Data Lake

Analytics Service

Microsoft

Azure Data Lake Analytics

All data



Productivity from day one

Easy and powerful data preparation

Limitless scale

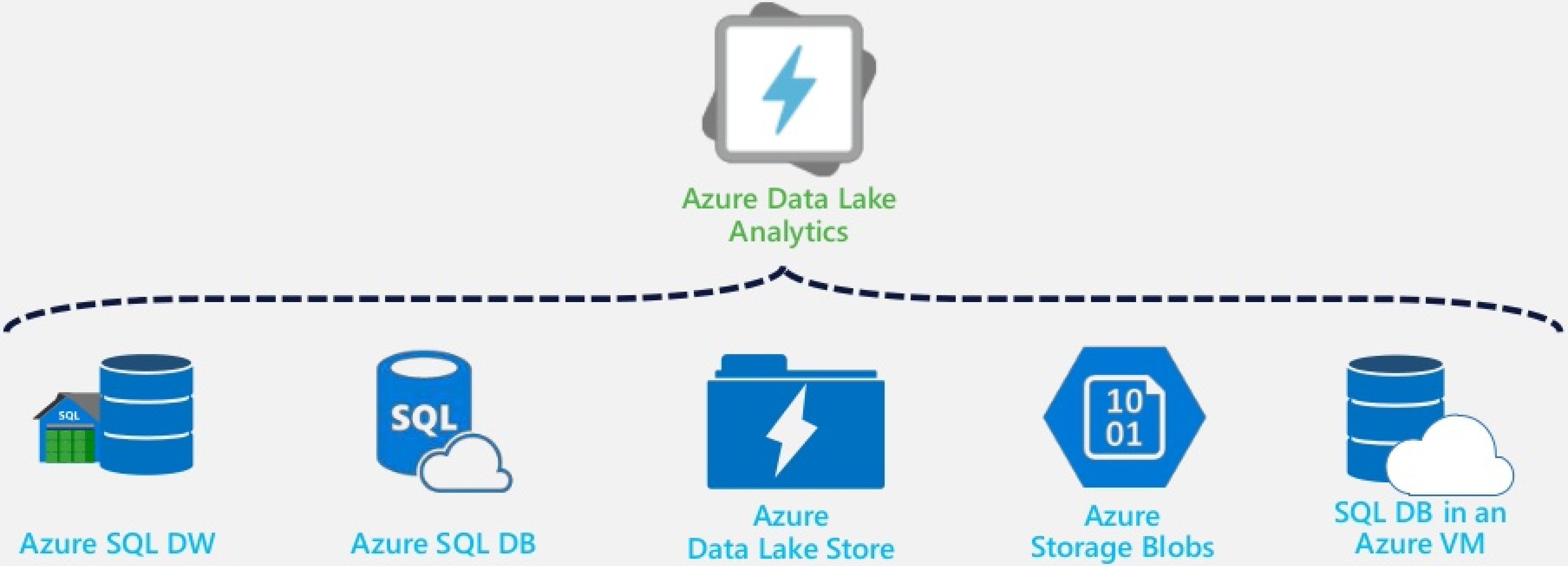
Enterprise-grade

Azure Data Lake Analytics Service

A new distributed analytics service

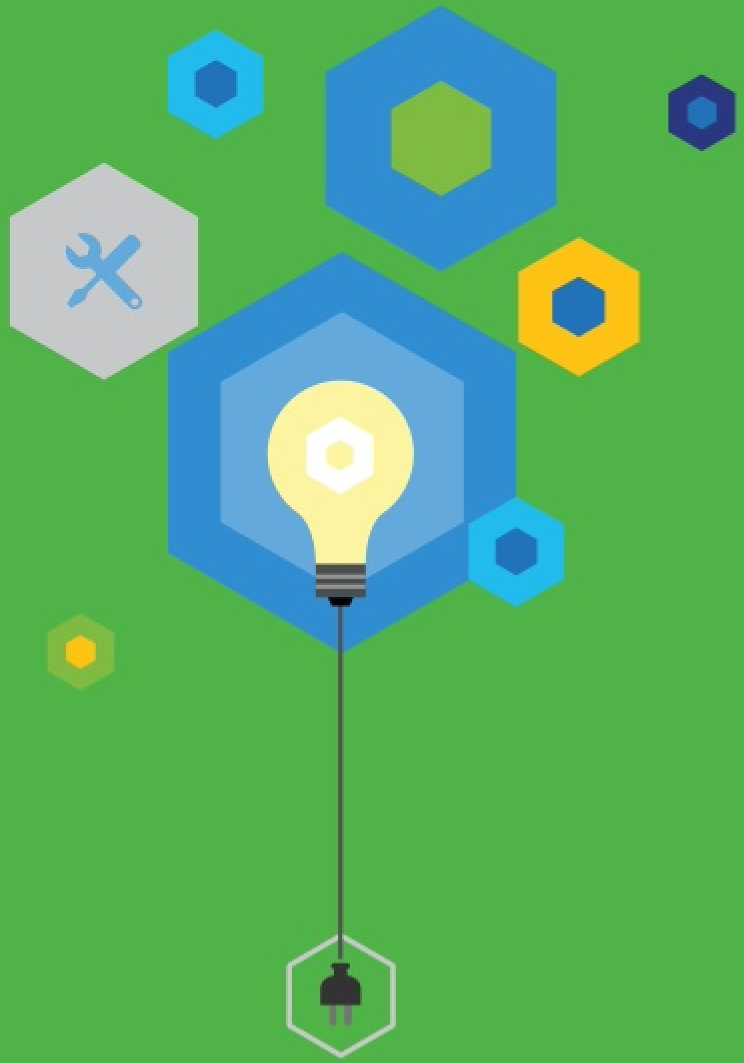
- Built on **Apache YARN**
- **Scales dynamically** with the turn of a dial
- **Pay by the query**
- Supports **Azure AD** for access control, roles, and integration with on-prem identity systems
- Built with **U-SQL** to unify the benefits of SQL with the power of C#
- Processes data **across Azure**

Work across all cloud data



Microsoft

Demo
Show me ADL!



Azure Data Lake

U-SQL

Microsoft

Why U-SQL?

Characteristics of Big Data Analytics

- Requires processing of any type of data
- Allow use of custom algorithms
- Scale to any size and be efficient

Some sample use cases

Digital Crime Unit – Analyze complex attack patterns to understand BotNets and to predict and mitigate future attacks by **analyzing log records with complex custom algorithms**

Image Processing – **Large-scale** image feature extraction and classification using **custom code**

Shopping Recommendation – Complex pattern analysis and prediction over shopping records using **proprietary algorithms**

Status Quo: SQL for Big Data

- ☺ **Declarativity does scaling and parallelization for you**
- ☹ **Extensibility is bolted on and not “native”**
  - ☹ hard to work with anything other than structured data
  - ☹ difficult to extend with custom code

Status Quo: Programming Languages for Big Data

- ☺ **Extensibility through custom code is "native"**
- ☹ **Declarativity is bolted on and not "native"**
  - ☹ User often has to care about scale and performance
  - ☹ SQL is 2nd class within string
  - ☹ Often no code reuse/ sharing across queries

Why U-SQL?

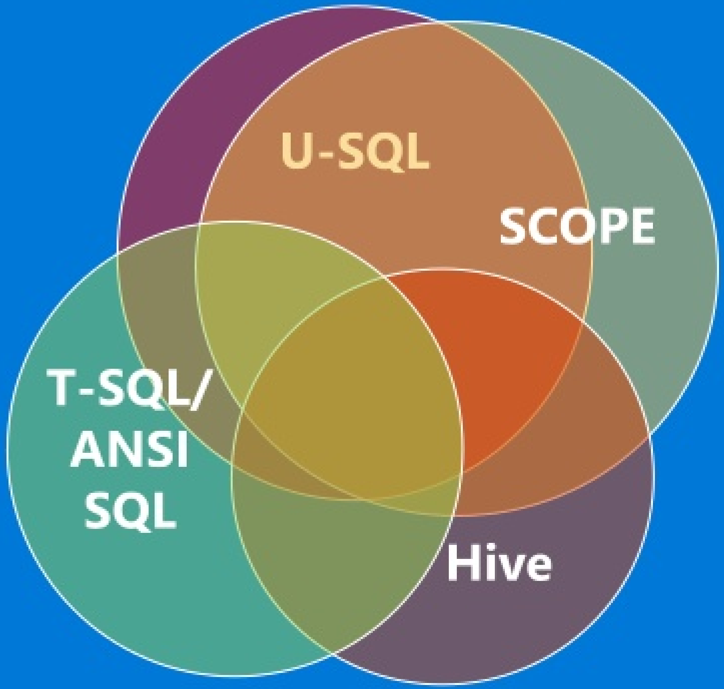
☺☺ **Declarativity and Extensibility are equally native to the language!**

Get benefits of both!

Makes it **easy** for you by **unifying**:

- Unstructured and structured data processing
- Declarative SQL and custom imperative Code
- Local and remote Queries
- Increase productivity and agility from Day 1 and at Day 100 for **YOU**!

The origins of U-SQL



SCOPE – Microsoft's internal Big Data language

- SQL and C# integration model
- Optimization and Scaling model
- Runs 100'000s of jobs daily

Hive

- Complex data types (Maps, Arrays)
- Data format alignment for text files

T-SQL/ANSI SQL

- Many of the SQL capabilities (windowing functions, meta data model etc.)

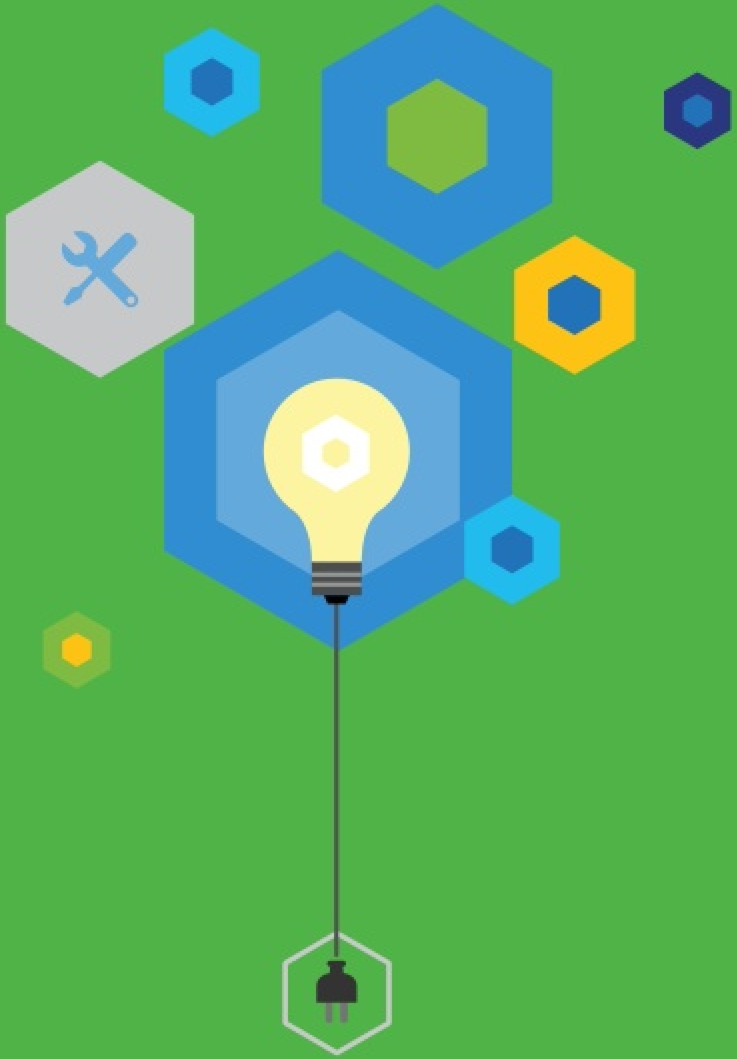
U-SQL extensibility

Extend U-SQL with C#/.NET

Built-in operators, function, aggregates

- C# expressions (in SELECT expressions)
- User-defined operators (UDOs)
- User-defined functions (UDFs)
- User-defined aggregates (UDAGGs)

Demo
Show me U-SQL!



Microsoft

U-SQL Language Philosophy

Declarative Query and Transformation Language:

- Uses SQL's SELECT FROM WHERE with GROUP BY/Aggregation, Joins, SQL Analytics functions
- Optimizable, Scalable

Expression-flow programming style:

- Easy to use functional lambda composition
- Composable, globally optimizable

Operates on Unstructured & Structured Data

- Schema on read over files
- Relational metadata objects (e.g. database, table)

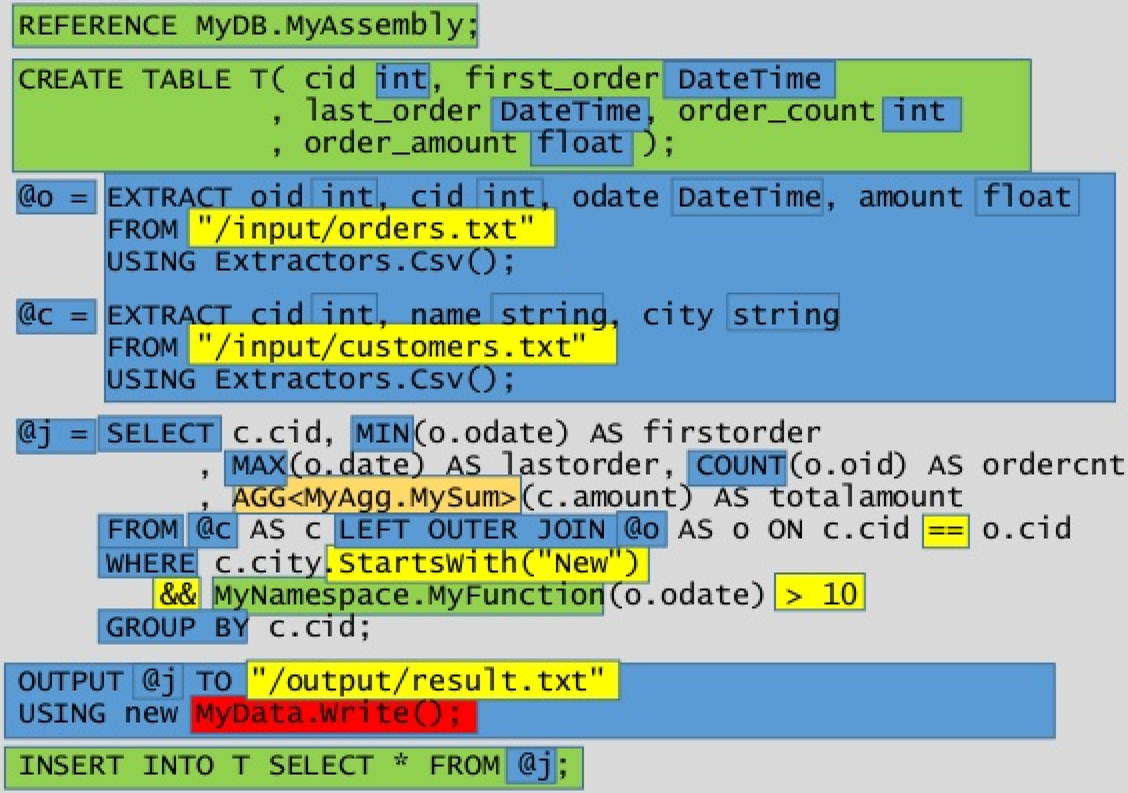
Extensible from ground up:

- Type system is based on C#
- Expression language IS C#
- User-defined functions (U-SQL and C#)
- User-defined Aggregators (C#)
- User-defined Operators (UDO) (C#)

U-SQL provides the Parallelization and Scale-out Framework for Usercode

- EXTRACTOR, OUTPUTTER, PROCESSOR, REDUCER, COMBINER, APPLIER

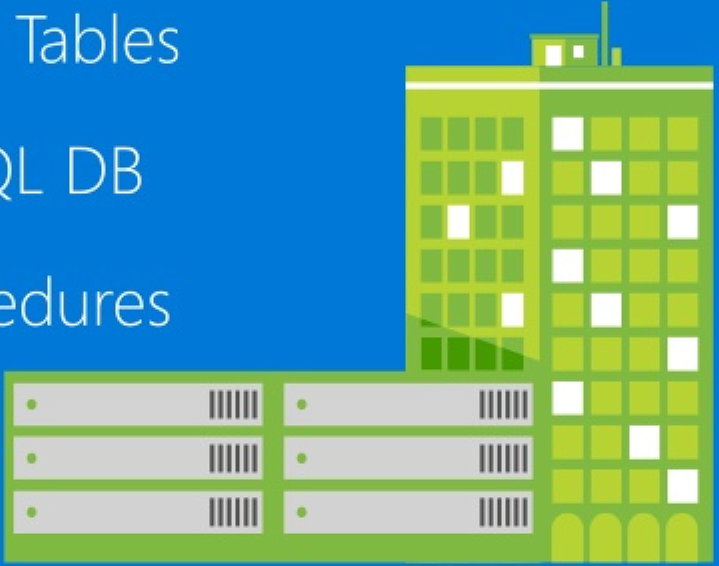
Federated query across distributed data sources



```
REFERENCE MyDB.MyAssembly;

CREATE TABLE T( cid int, first_order DateTime
              , last_order DateTime, order_count int
              , order_amount float );

@o = EXTRACT oid int, cid int, odate DateTime, amount float
     FROM "/input/orders.txt"
     USING Extractors.Csv();

@c = EXTRACT cid int, name string, city string
     FROM "/input/customers.txt"
     USING Extractors.Csv();

@j = SELECT c.cid, MIN(o.odate) AS firstorder
          , MAX(o.date) AS lastorder, COUNT(o.oid) AS ordercnt
          , AGG<MyAgg.MySum>(c.amount) AS totalamount
     FROM @c AS c LEFT OUTER JOIN @o AS o ON c.cid == o.cid
     WHERE c.city.StartsWith("New")
        && MyNamespace.MyFunction(o.odate) > 10
     GROUP BY c.cid;

OUTPUT @j TO "/output/result.txt"
USING new MyData.Write();

INSERT INTO T SELECT * FROM @j;
```

Lots of additional interesting features

- ✓ Set of files with patterns, Partitioned Tables
- ✓ Federated Queries against Azure SQL DB
- ✓ Views, Table-Valued Functions, Procedures
- ✓ SQL Windowing Functions
- ✓ Complex Types (MAP, ARRAY)



Intro Blog entry: http://aka.ms/usql-intro
Blog entry on UDFs: http://aka.ms/usql-udf
U-SQL Reference Doc (beta): http://aka.ms/usql_reference
U-SQL Community & Team site: http://usql.io/
Videos: https://channel9.msdn.com/Series/AzureDataLake

Additional Resources

- Blogs and community page:
  - http://usql.io
  - https://blogs.msdn.microsoft.com/azuredatalake/
  - http://blogs.msdn.com/b/visualstudio/
  - http://azure.microsoft.com/en-us/blog/topics/big-data/
  - https://channel9.msdn.com/Search?term=U-SQL#ch9Search
- Documentation:
  - http://aka.ms/usql_reference
  - https://azure.microsoft.com/en-us/documentation/services/data-lake-analytics/
- ADL forums and feedback
  - http://aka.ms/adlfeedback
  - https://social.msdn.microsoft.com/Forums/azure/en-US/home?forum=AzureDataLake
  - http://stackoverflow.com/questions/tagged/u-sql

This is why U-SQL!

- ✓ Unifies natively SQL's declarativity and C#'s extensibility
- ✓ Unifies querying structured and unstructured
- ✓ Unifies local and remote queries
- ✓ Increase productivity and agility from Day 1 forward for YOU!

→ Sign up for an Azure Data Lake account and join the Public Preview http://www.azure.com/datalake and give us your feedback via http://aka.ms/adlfeedback or at http://aka.ms/u-sql-survey!