Azure Data Lake

Marek Chmel | MVP: Data Platform, Principal Architect

marek@datascienceteam.cz
@MarekChmel
Key Message

With the increase of computing power, electronic devices and accessibility to the Internet, more data than ever is being produced, collected and transmitted.

Organizations have recognized the power of data analysis, but are struggling to manage the massive amounts of information they have.
Every day we create
2,500,000,000,000,000
(2.5 quintillion) bytes of data

90% of the world’s data today has been created in the last 2 years alone.

This would fill 10 million Blu-ray discs, the height of which stacked, would measure the height of 4 Eiffel Towers on top of one another.
BIG DATA:

- Data stored grows 4x faster than world economy

- Substantial shift in economic power and source of economic value

- Increasing quantity of data allows for more qualitative approach
Introducing Data Lakes
Two approaches to information management for analytics

Top-down (deductive)
- Theory
- Hypothesis
- Observation
- Confirmation

Bottom-up (inductive)
- Theory
- Hypothesis
- Pattern
- Observation

Two approaches to information management for analytics:

1. **Bottom-up (inductive)**
   - Observation
   - Pattern
   - Hypothesis
   - Theory

   - What happened?
   - Why did it happen?
   - What will happen?
   - How can we make it happen?

   - Descriptive analytics
   - Diagnostic analytics
   - Predictive analytics
   - Prescriptive analytics

2. **Top-down (deductive)**
   - Theory
   - Hypothesis
   - Observation
   - Confirmation

   - What happened?
   - Why did it happen?
   - What will happen?
   - How can we make it happen?

   - Descriptive analytics
   - Diagnostic analytics
   - Predictive analytics
   - Prescriptive analytics
The data lake uses a bottom-up approach

Ingest all data regardless of requirements

Store all data in native format without schema definition

Do analysis using analytic engines like Hadoop

Batch queries
Interactive queries
Real-time analytics
Machine Learning
Data warehouse
Data warehousing uses a top-down approach

Understand corporate strategy

Gather requirements

Implement data warehouse

Reporting and analytics design

Dimension modeling

ETL design

Set up infrastructure

Reporting and analytics development

Physical design

ETL development

Install and tune

Business requirements

Technical requirements

Data sources

BI and analytics

Data warehouse

ETL

OLTP

ERP

CRM

LOB
Options for Big Data in Azure

Clusters

Virtual Machines
User-managed VM cluster running Hadoop

HDInsight
Azure-managed VM cluster running Hadoop

Serverless

Data Lake
Big Data as a Service
- Data Lake Analytics
- Data Lake Storage Gen1
Challenges involved in implementing a data lake

**Data Silos**
- Data spans sources
- Inefficiency in colocation

**Analytics**
- Open interfaces to data
- Variety of analytics tools

**Performance and Scale**
- Storage bottlenecks
- IoT sources – small writes
- Price-performance
- Data grows independently

**Security**
- Compliance challenges
- Effectively control access
- Corporate policies
Azure Data Lake

Azure Data Lake (ADL)

Analytics

Azure Data Lake Analytics

Storage

Azure Data Lake Store
Azure Data Lake demystified

Azure Data Lake

---

**Data Lake Storage Gen1**

- Petabyte size files and trillions of objects
- Provide I/O capacity to bandwidth hungry apps like HDI, Cloudera and Hortonworks
- Encryption on disk
- File & Folder ACLs

---

**Data Lake Analytics**

- Start in seconds, scale instantly, pay per job
- Develop massively parallel programs with simplicity
- Leverage open source (Hadoop, Spark, Python, R...)
- Debug & optimize big data programs with ease
- Virtualize your analytics

---

- Always encrypted
- Integrated Azure Active Directory
- Role-based access Control
- Enterprise-grade support
Built on open source

- ADL Analytics
  - U/SQL
- ADL HDInsight
  - Hive
  - Spark
  - STORM
- YARN
- WebHDFS
- ADL Store
Data lakes in Azure Cloud

On-premises

Hadoop cluster
- MAP
- Reduce
- Hive, Pig
- HBase
- Storm
- YARN-based Compute
- HDFS/WebHDFS API
  - Hadoop Distributed File System

Azure Cloud

Azure HDInsight
- MAP
- Reduce
- Hive, Pig
- HBase
- Storm
- YARN-based Compute

Azure Data Lake
- WebHDFS API
  - Hadoop Distributed File System
Cortana Analytics is a fully managed big data and advanced analytics suite that enables you to transform your data into intelligent action.
Big data made easy

Analytics on any data, any size

All users productive on day one

Ready for your enterprise
Understanding ADL Store
What is Azure Data Lake (ADL) Store?
Comparing ADLS with Azure Blob Storage

Azure Data Lake Store
- Optimized for analytics

Azure Blob storage
- Bulk storage of files
- Cold data storage
How do you start using ADL?

1. Log in to the Azure portal.
2. Create an ADL Analytics account (90 seconds, free).
3. Write a U-SQL script and submit it to the ADL Analytics account.
4. The U-SQL job reads and writes data.
DEMO

Azure Data Lake Store
Understanding ADLA
Azure Data Lake Analytics service

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

Supports Azure Active Directory

Access control
Roles
Integration with on-premises identity systems

Includes U-SQL

Benefits of SQL
Expressive power of C#

Processes data across multiple Azure data sources

U-SQL scalable runtime
Simplified management and administration

- Web-based management in Azure portal
- Automate tasks using PowerShell
- Role-based Access Control with Azure Active Directory
- Monitor service operations and activity
DEMO

Azure Data Lake Analytics
What is U-SQL?

U-SQL is a hyper-scalable, highly extensible language for preparing, transforming, and analyzing all data.

It allows users to focus on the what—not the how—of business problems.

Built on familiar languages (SQL and C#, SCOPE) and supported by a fully integrated development environment.

Built for data developers and scientists.
U-SQL fundamentals

All the familiar SQL clauses

Operate on unstructured and structured data

Relational metadata objects
U-SQL queries: General pattern

Read

- Azure Storage blobs
- Azure SQL DB
- Azure Data Lake

Process

- EXTRACT
- SELECT
- RowSet
- SELECT ...
- FROM...
- WHERE ...
- RowSet

Store

- OUTPUT
- INSERT
- Azure Storage blobs
- Azure Data Lake
.NET integration and extensibility

U-SQL expressions are full C# expressions

Reuse .NET code in your own assemblies

Use C# to define your own:

- Types
- Functions
- Joins
- Aggregators
- IO (Extractors, Outputters)
### Query across Azure data sources

<table>
<thead>
<tr>
<th></th>
<th>READ</th>
<th>WRITE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Azure Data Lake Store</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Azure Storage blobs</td>
<td>✅</td>
<td>✅</td>
</tr>
<tr>
<td>Azure SQL database</td>
<td>✅</td>
<td></td>
</tr>
<tr>
<td>Azure SQL data warehouse</td>
<td>✅</td>
<td></td>
</tr>
<tr>
<td>Azure SQL DB in Azure VM</td>
<td>✅</td>
<td></td>
</tr>
</tbody>
</table>
## Visual Studio and U-SQL integration

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author U-SQL scripts (with C# code)</td>
<td>Debug U-SQL and C# code</td>
</tr>
<tr>
<td>Visualize and replay progress of job</td>
<td>Fine-tune query performance</td>
</tr>
<tr>
<td>Submit and cancel U-SQL jobs</td>
<td>Create metadata objects</td>
</tr>
<tr>
<td>Visualize physical plan of U-SQL query</td>
<td>Browse metadata catalog</td>
</tr>
</tbody>
</table>

*U-SQL* stands for Universal SQL, a language designed for data analytics that integrates seamlessly with Visual Studio.
Read the input, write it directly to output (just a simple copy)

@rows =
EXTRACT
    OrderId  int,
    Customer string,
    Date    DateTime,
    Amount  float
FROM "/input/orders.txt"
USING Extractor.Tsv();

OUTPUT @rows
    TO "adl://mylake/orders_copy.txt"
USING Outputters.Tsv();
Follow up

- Course 20775A: Performing Data Engineering on Microsoft HD Insight
- Course 20774A: Perform Cloud Data Science with Azure Machine Learning
- Course 20776A: Performing Big Data Engineering on Microsoft Cloud Services