Azure Data Lake Store
Azure Data Lake Analytics

A technical overview and introduction to U-SQL

Gary Hope
Cloud Data Solution Architect
Microsoft South Africa
GaryHope@Microsoft.com

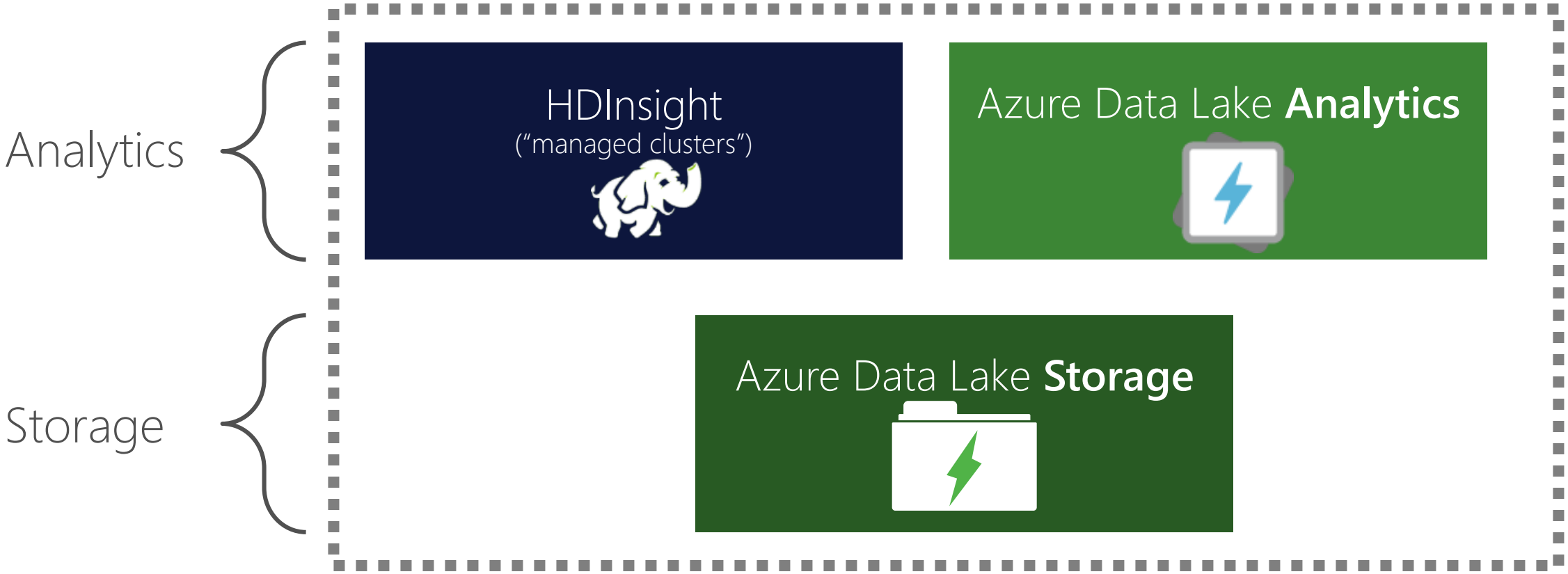
Microsoft

PASS SQLSATURDAY
JOHANNESBURG | 02 SEP 2017

Proudly brought to you by

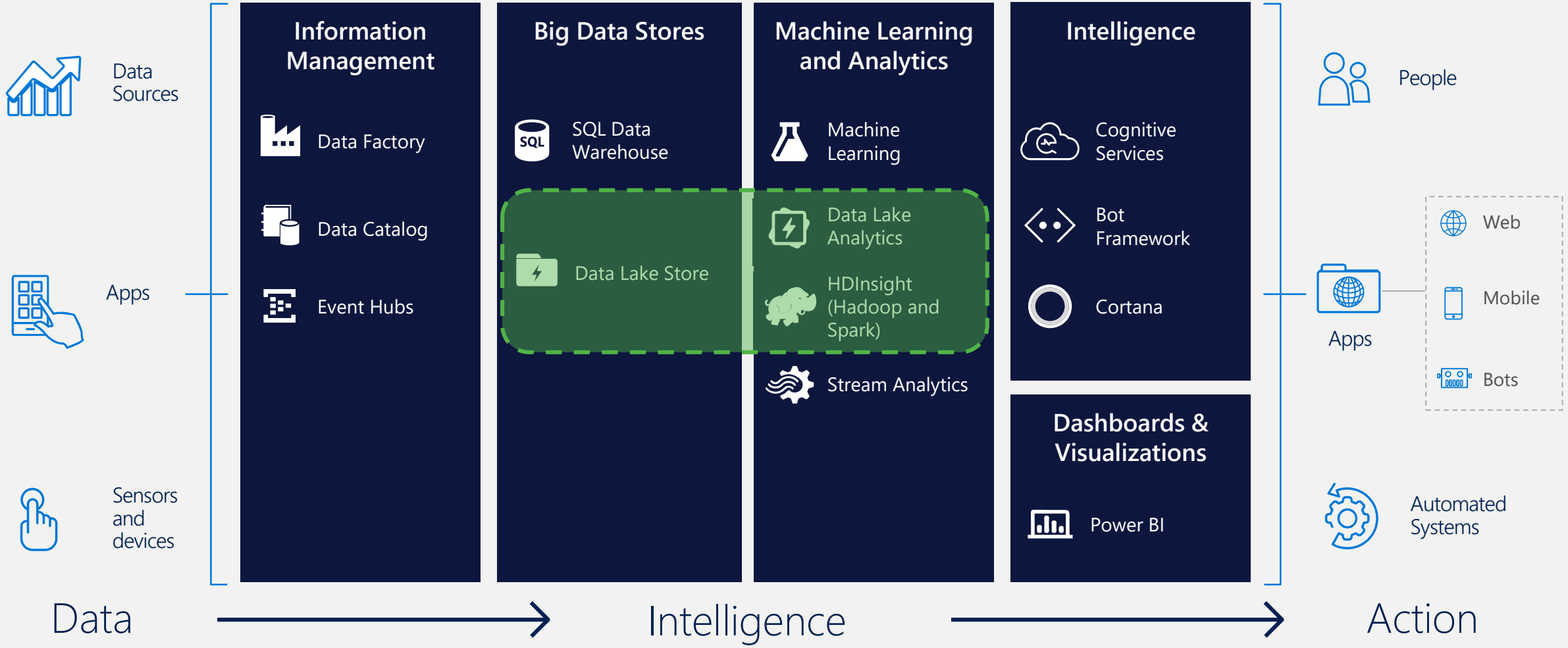
Platinum	Microsoft
Gold	EOH, First Technology Enterprise Data Solutions (Pty) Ltd
Silver	entelect everything is possible, KARABINA™
Bronze	ENTERPRISEWORX

Azure Data Lake



Azure Data Lake

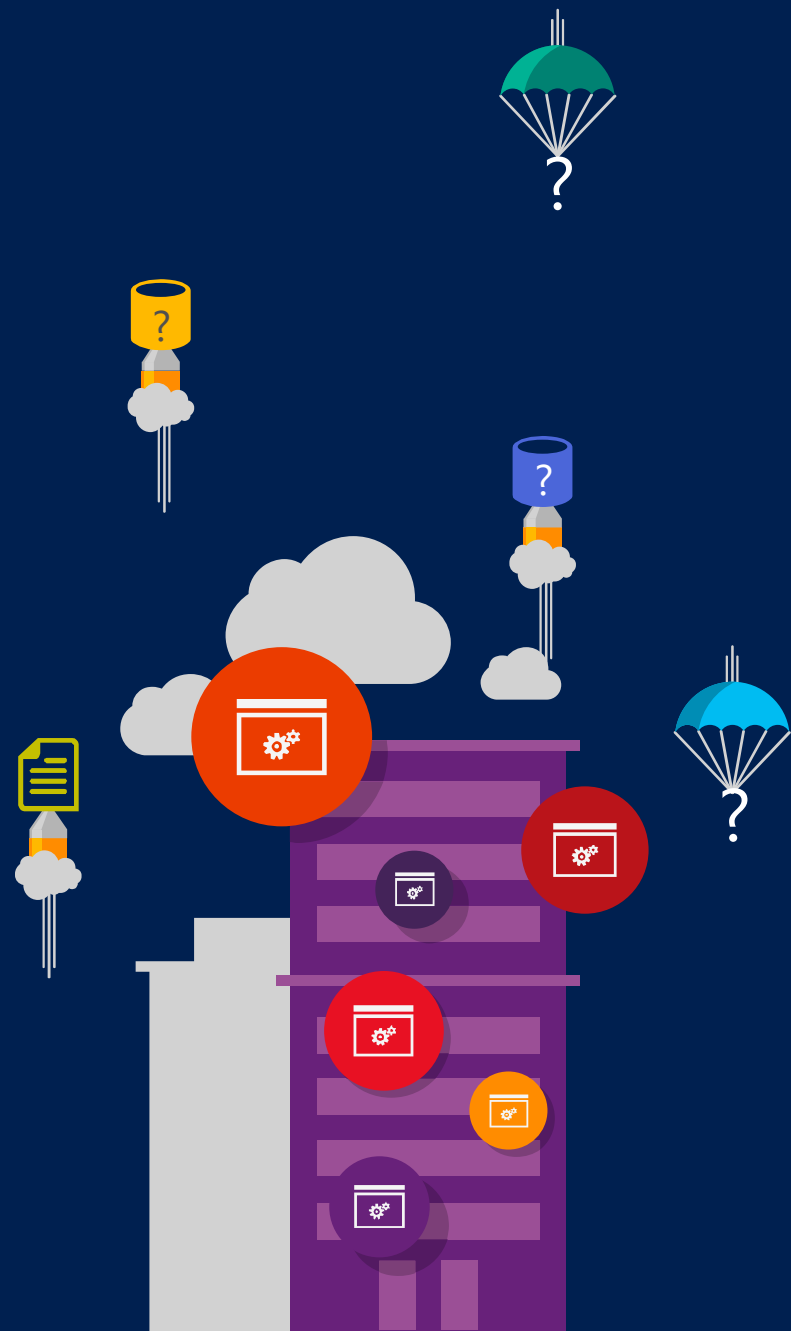
as part of Cortana Intelligence Suite



Azure Data Lake

Demo – Lets Create The Services

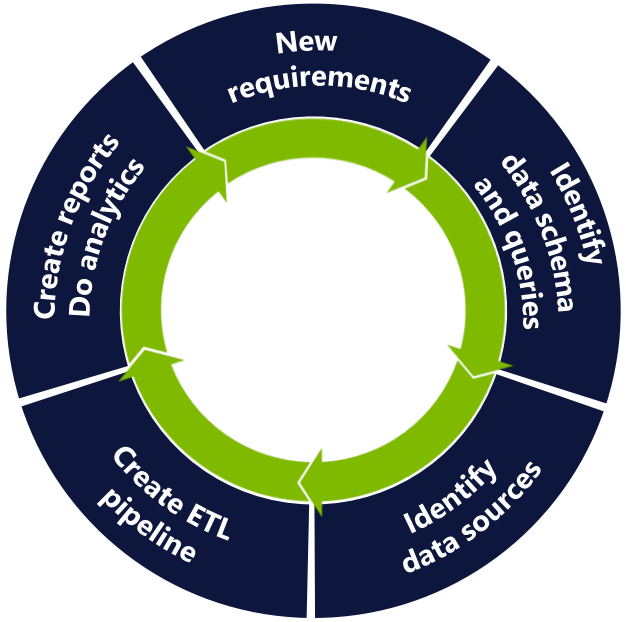
Why data lakes?

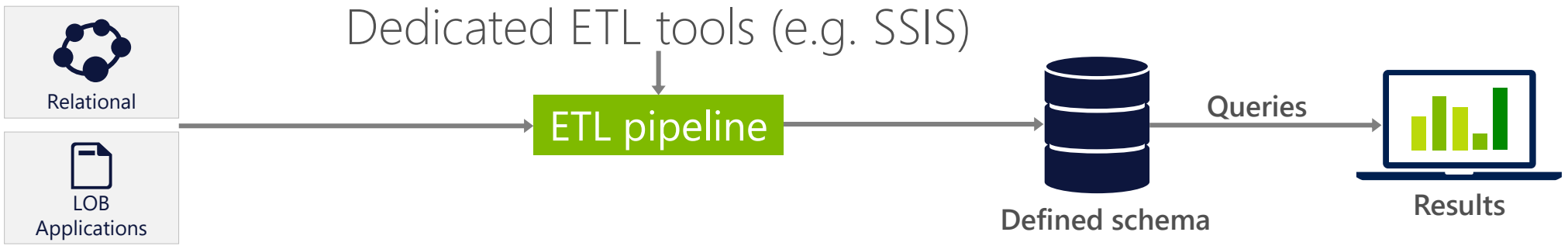


Microsoft

Traditional business analytics process

1. Start with end-user requirements to identify desired reports and analysis
2. Define corresponding database schema and queries
3. Identify the required data sources
4. Create a Extract-Transform-Load (ETL) pipeline to extract required data (curation) and transform it to target schema (*'schema-on-write'*)
5. Create reports. Analyze data

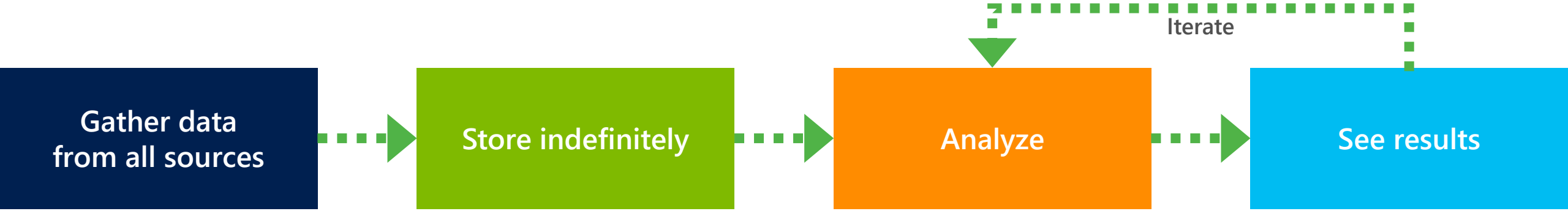




All data not immediately required is discarded or archived

New big data thinking: All data has value

- All data has potential value
- Data hoarding
- No defined schema—stored in native format
- Schema is imposed and transformations are done at query time *(schema-on-read)*.
- Apps and users interpret the data as they see fit

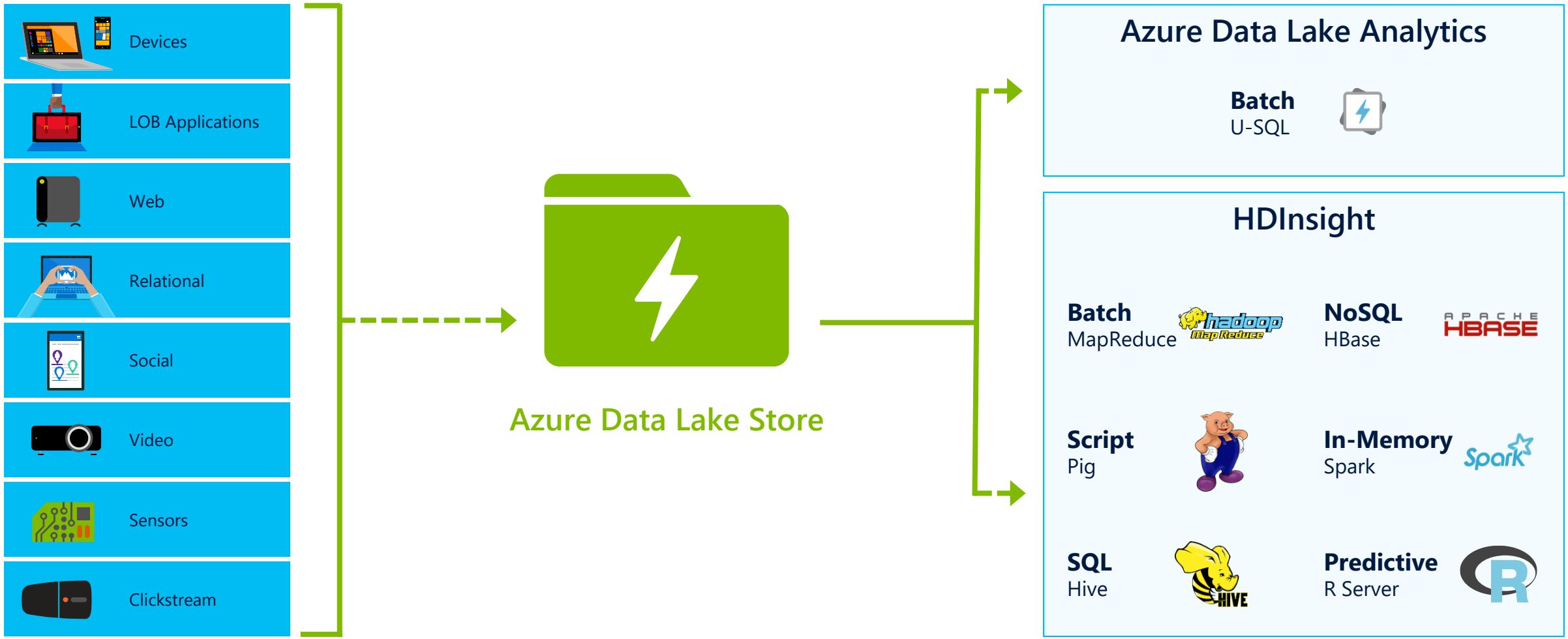


Data Lake Store: Technical Requirements

Secure	Must be highly secure to prevent unauthorized access (especially as all data is in one place).
Scalable	Must be highly scalable. When storing all data indefinitely, data volumes can quickly add up
Reliable	Must be highly available and reliable (no permanent loss of data).
Throughput	Must have high throughput for massively parallel processing via frameworks such as Hadoop and Spark
Details	Must be able to store data with all details; aggregation may lead to loss of details.
Native format	Must permit data to be stored in its 'native format' to track lineage & for data provenance.
All sources	Must be able ingest data from a variety of sources-LOB/ERP, Logs, Devices, Social NWs etc.
Multiple analytic frameworks	Must support multiple analytic frameworks—Batch, Real-time, Streaming, ML etc. No one analytic framework can work for all data and all types of analysis.

Big Data analytics workloads

A highly scalable, distributed, parallel file system in the cloud specifically designed to work with a variety of big data analytics workloads



Azure Data Lake Store

Scale, Performance, Reliability

Microsoft

Azure Data Lake Store: No Scale Limits

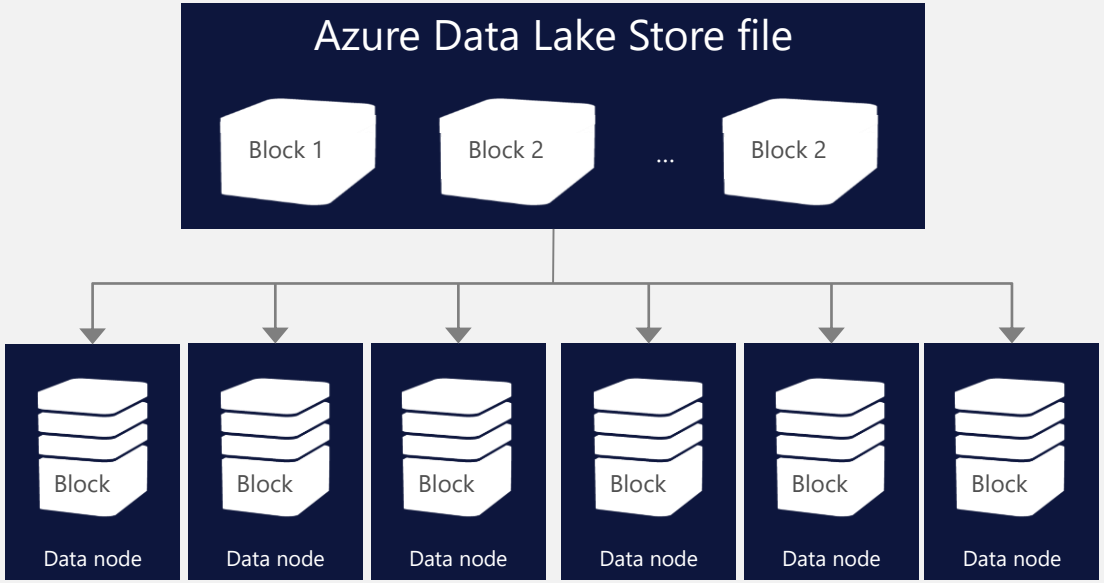
Azure Data Lake Store integrates with Azure Active Directory (AAD) for:

- Amount of data stored
- How long data can be stored
- Number of files
- Size of the individual files
- Ingestion throughput

Seamlessly scales from a few KBs to several PBs

Azure Data Lake Store: How it works

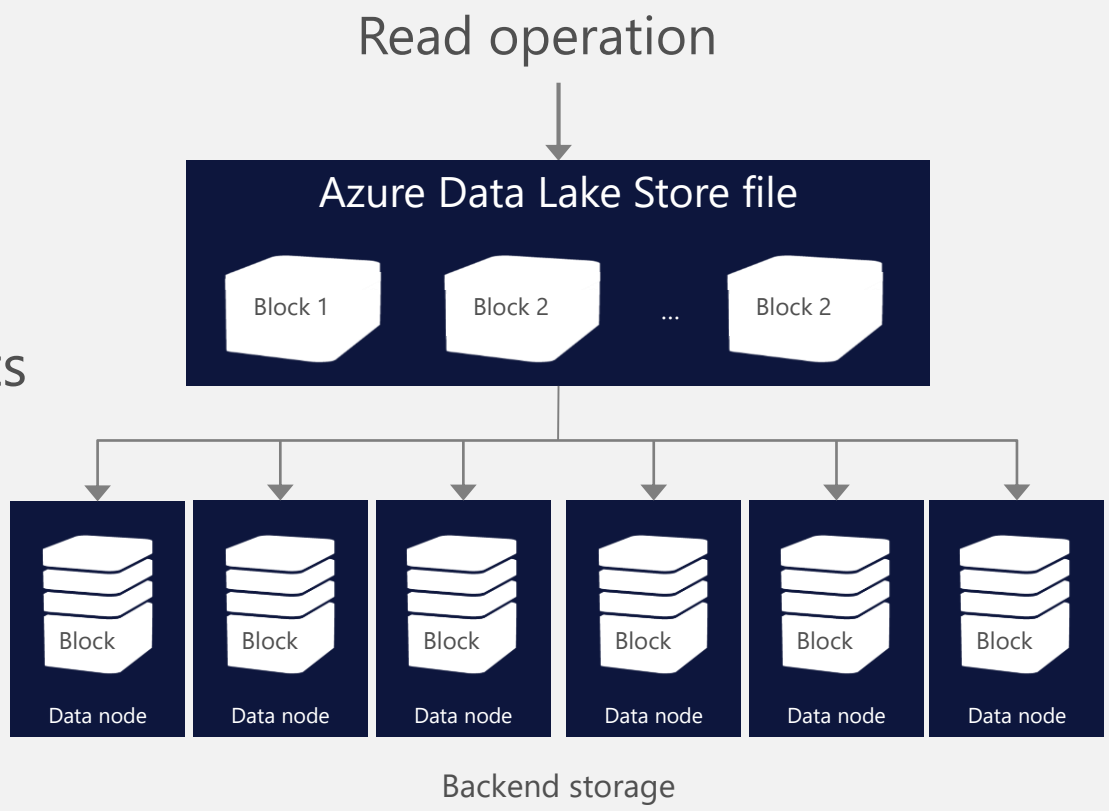
- Each file in ADL Store is sliced into blocks
- Blocks are distributed across multiple data nodes in the backend storage system
- With sufficient number of backend storage data nodes, files of any size can be stored
- Backend storage runs in the Azure cloud which has virtually unlimited resources
- Metadata is stored about each file
  No limit to metadata either.



Backend Storage

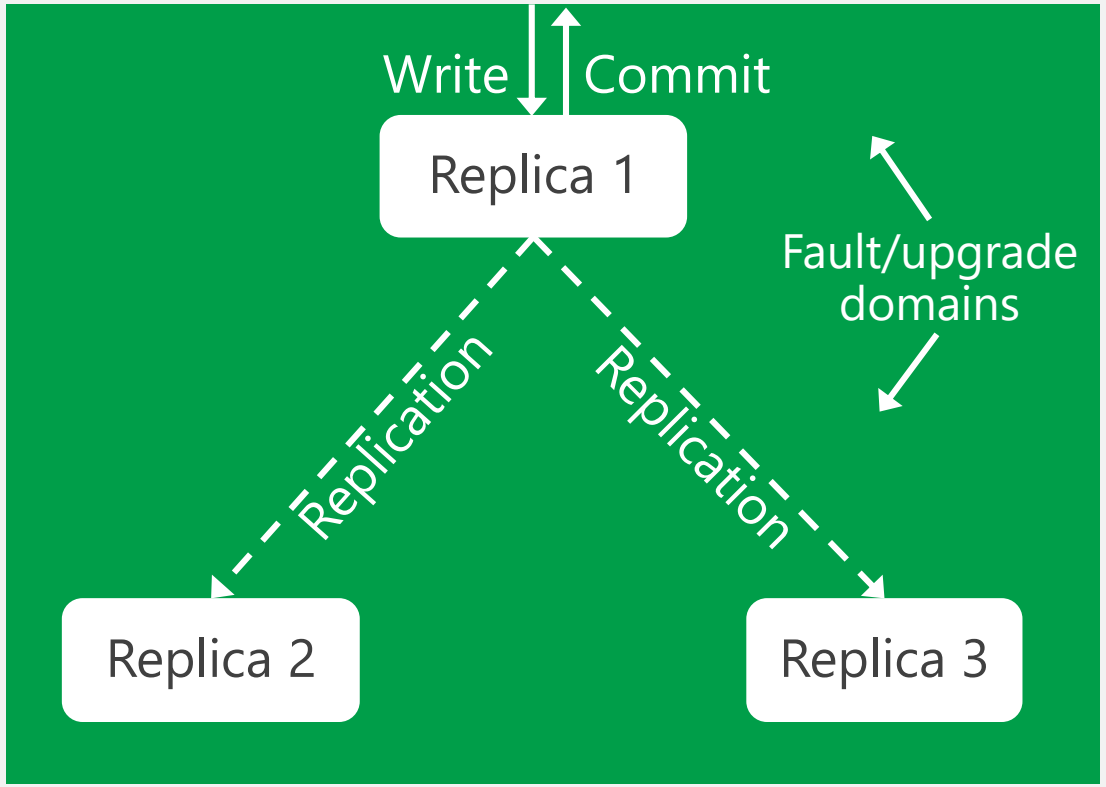
Azure Data Lake Store: Massive throughput

- Through read parallelism ADL Store provides massive throughput
- Each read operation on a ADL Store file results in multiple read operations executed in parallel against the backend storage data nodes



ADL Store: High Availability and Reliability

- Azure maintains 3 replicas of each data object per region across three fault and upgrade domains
- Each create or append operation on a replica is replicated to other two
- Writes are committed to application only after all replicas are successfully updated
- Read operations can go against any replica



Data is never lost or unavailable even under failures

Azure Data Lake Store

Security

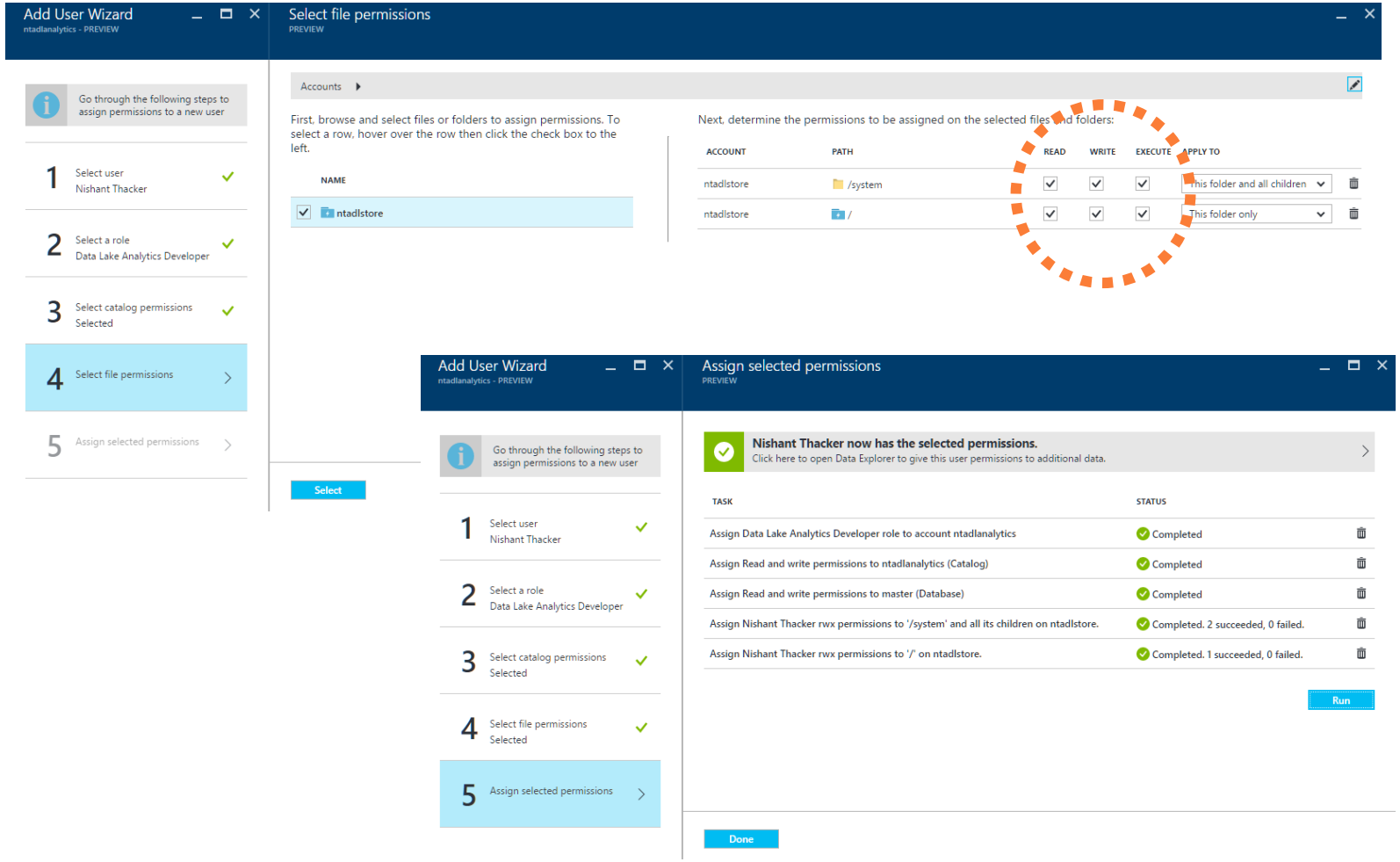
Microsoft

Azure Data Lake Store Security: AAD integration

- Multi-factor authentication based on OAuth2.0
- Integration with on-premises AD for federated authentication
- Role-based access control
- Privileged account management
- Application usage monitoring and rich auditing
- Security monitoring and alerting
- Fine-grained ACLs for AD identities

Azure Data Lake Store Security: Role-based access

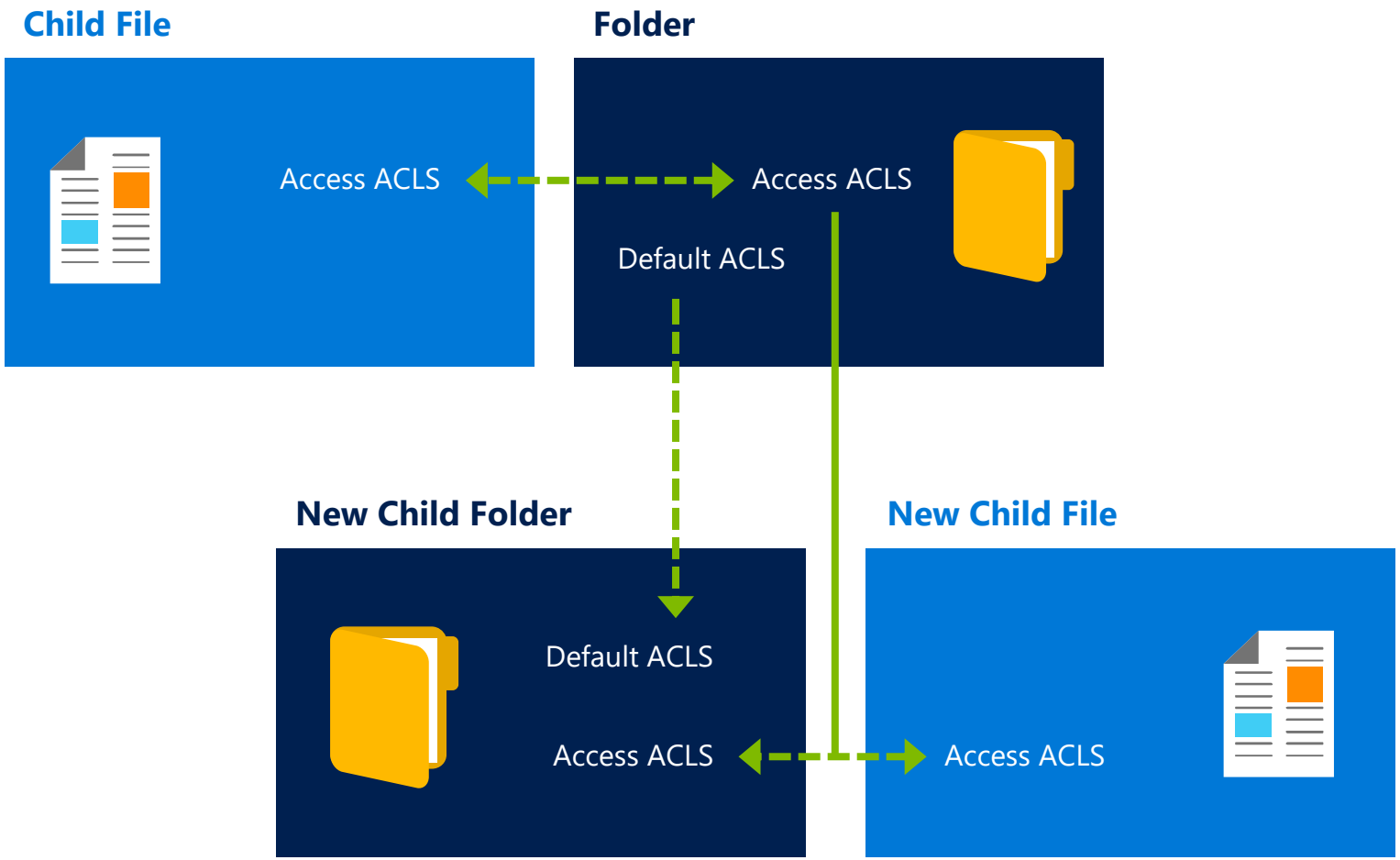
- Each file and directory is associated with an owner and a group
- Files or directories have separate permissions (read(r), write(w), execute(x)) for owners, members of the group, and for all other users
- Fine-grained access control lists (ACLs) rules can be specified for specific named users or named groups



Granular control of file and folder access

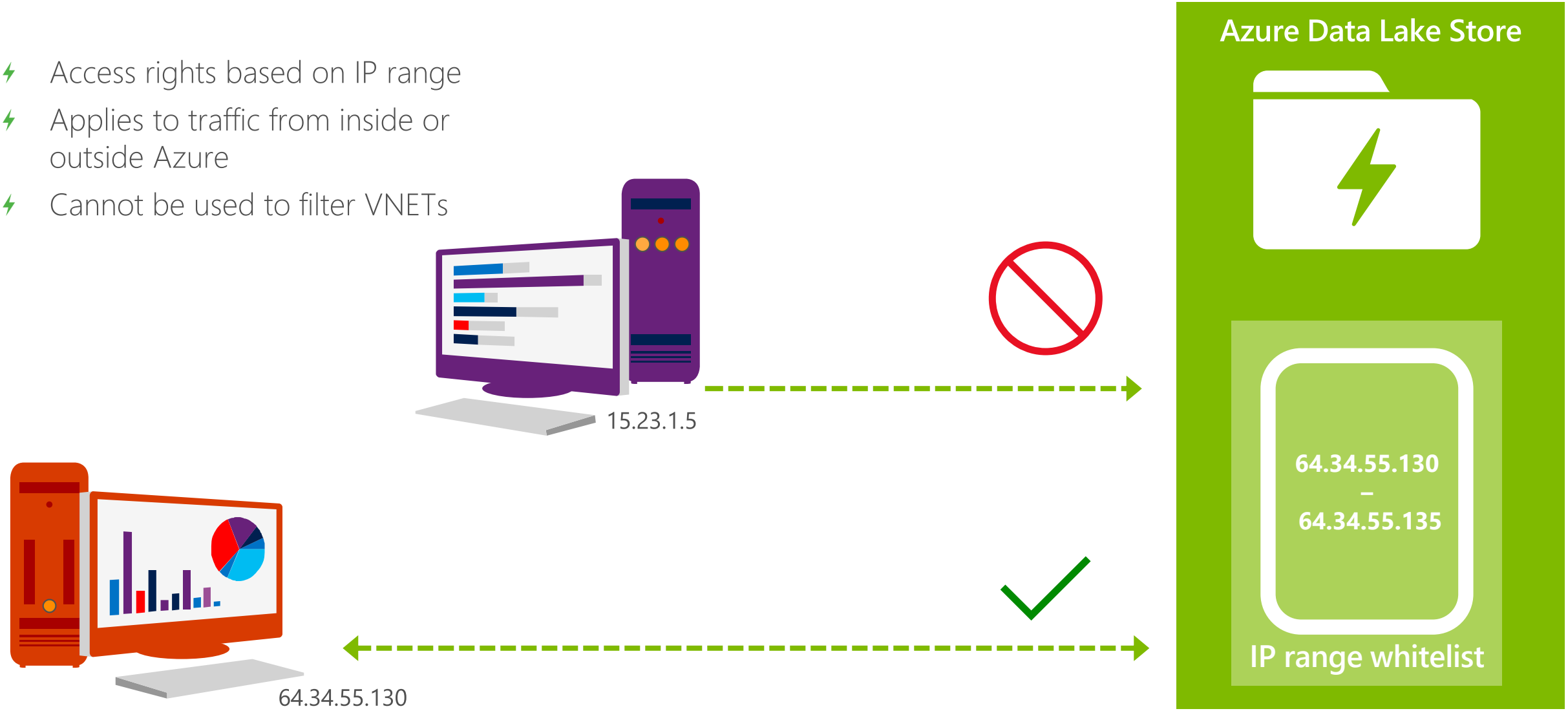
POSIX-Style ACLs with full compatibility with HDFS/WebHDFS

- Generate default ACLs for files and folders
- Customize for fine-tuned control
- Access ACLs control how a user can access to the file or folder
- Default ACLs used to construct the Access ACL of new children
- Default ACLs copied to the Default ACL of new child folders



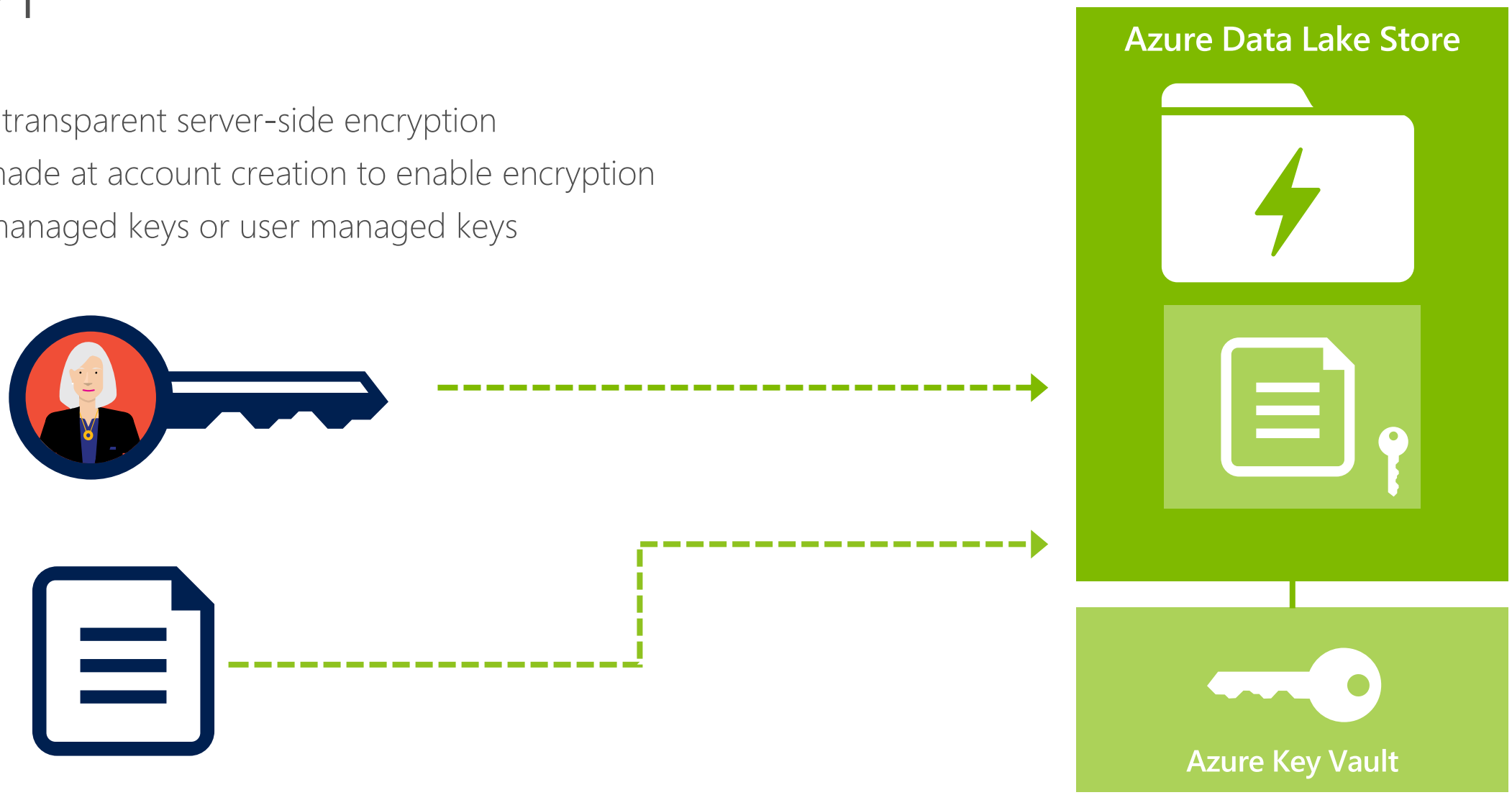
IP address ACLs

- Access rights based on IP range
- Applies to traffic from inside or outside Azure
- Cannot be used to filter VNETs



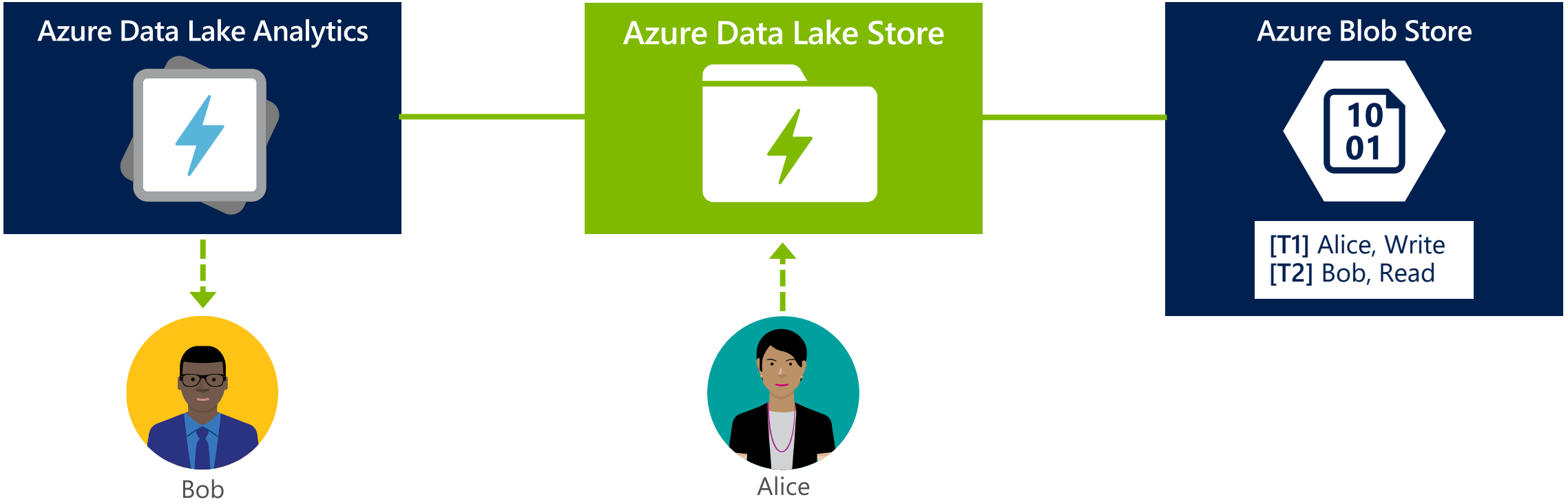
Encryption of data at rest

- Provides transparent server-side encryption
- Choice made at account creation to enable encryption
- Service managed keys or user managed keys



Audit logs for data access

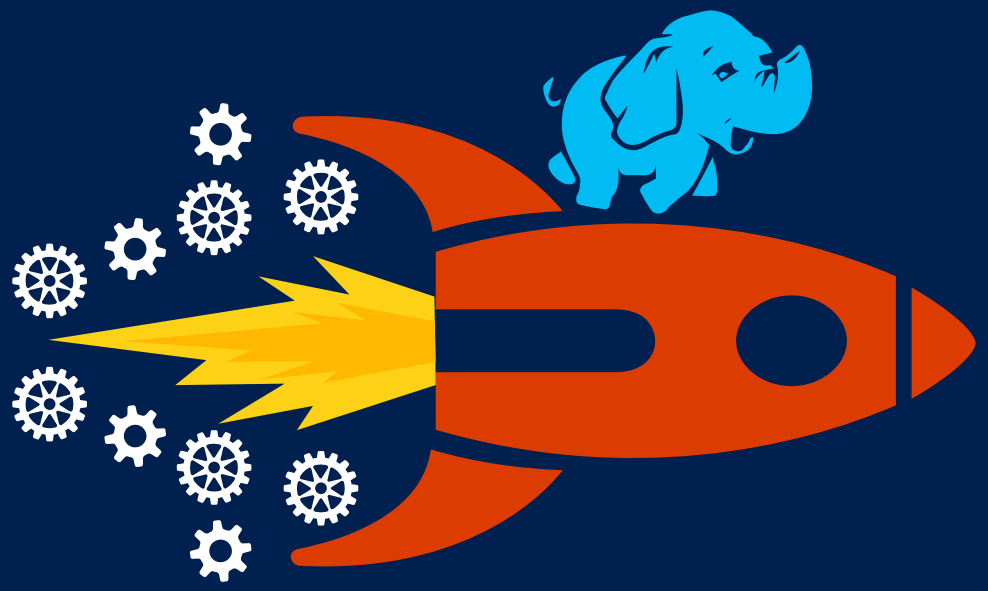
- Logs are available in JSON format
- Sample U-SQL scripts are available on GitHub to-read logs
- Enhancement to logs will continue through GA



Demo – Lets Upload Some Data

Azure Data Lake Store

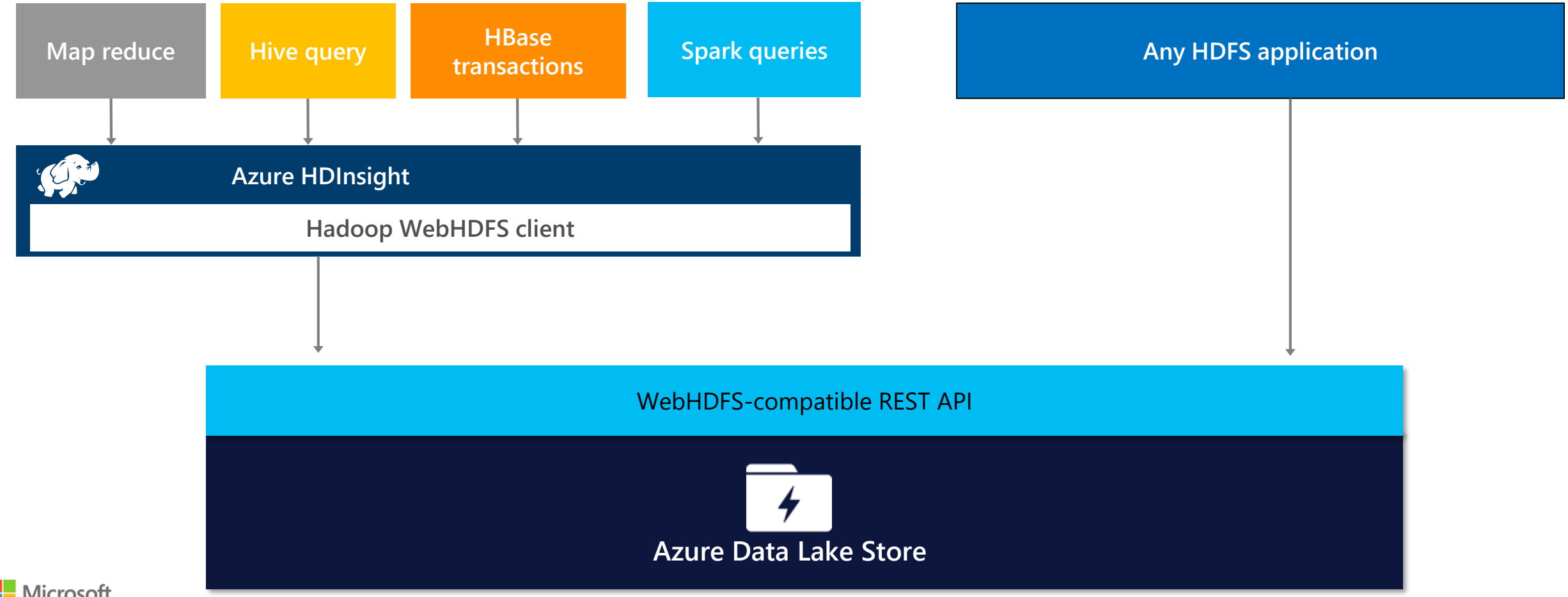
Hadoop integration and Data Movement



Microsoft

Azure Data Lake Store is HDFS-compatible

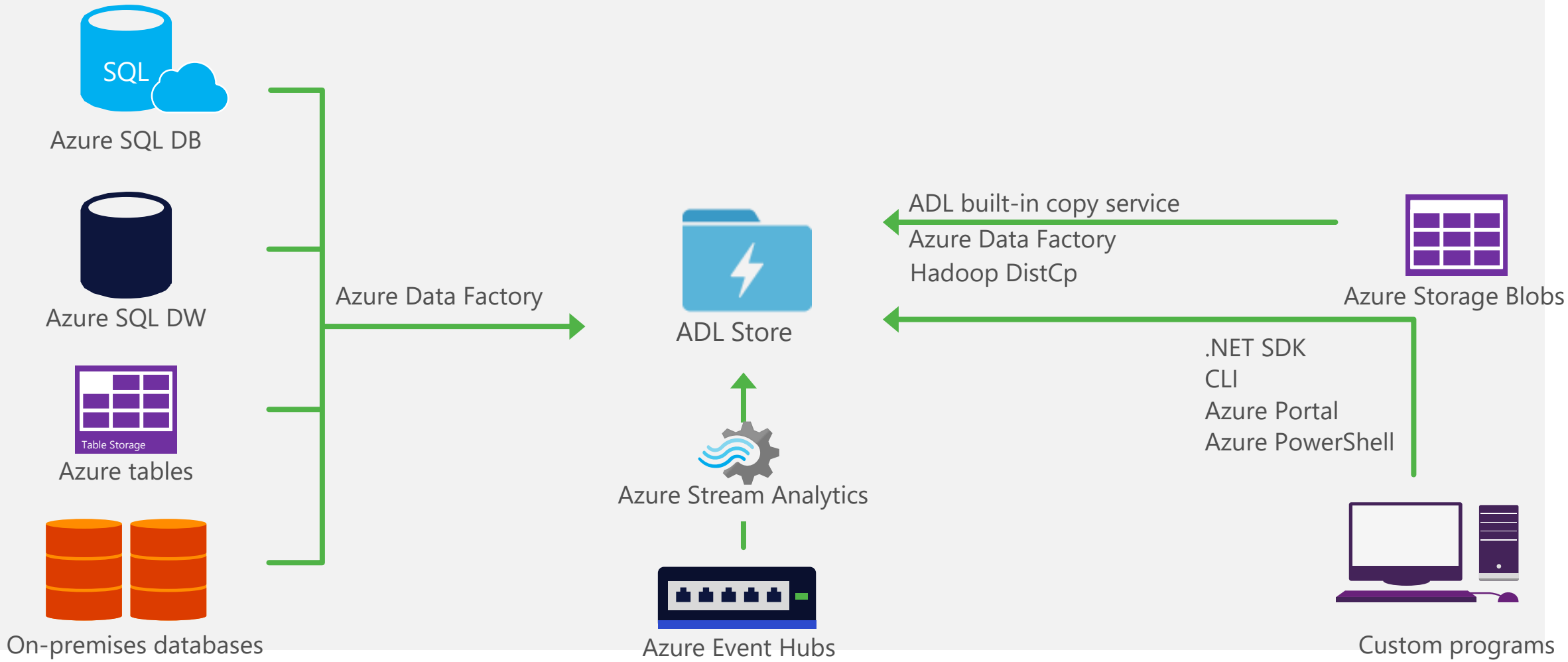
With a WebHDFS endpoint Azure Data Lake Store is a Hadoop-compatible file system that integrates seamlessly with Azure HDInsight



Microsoft

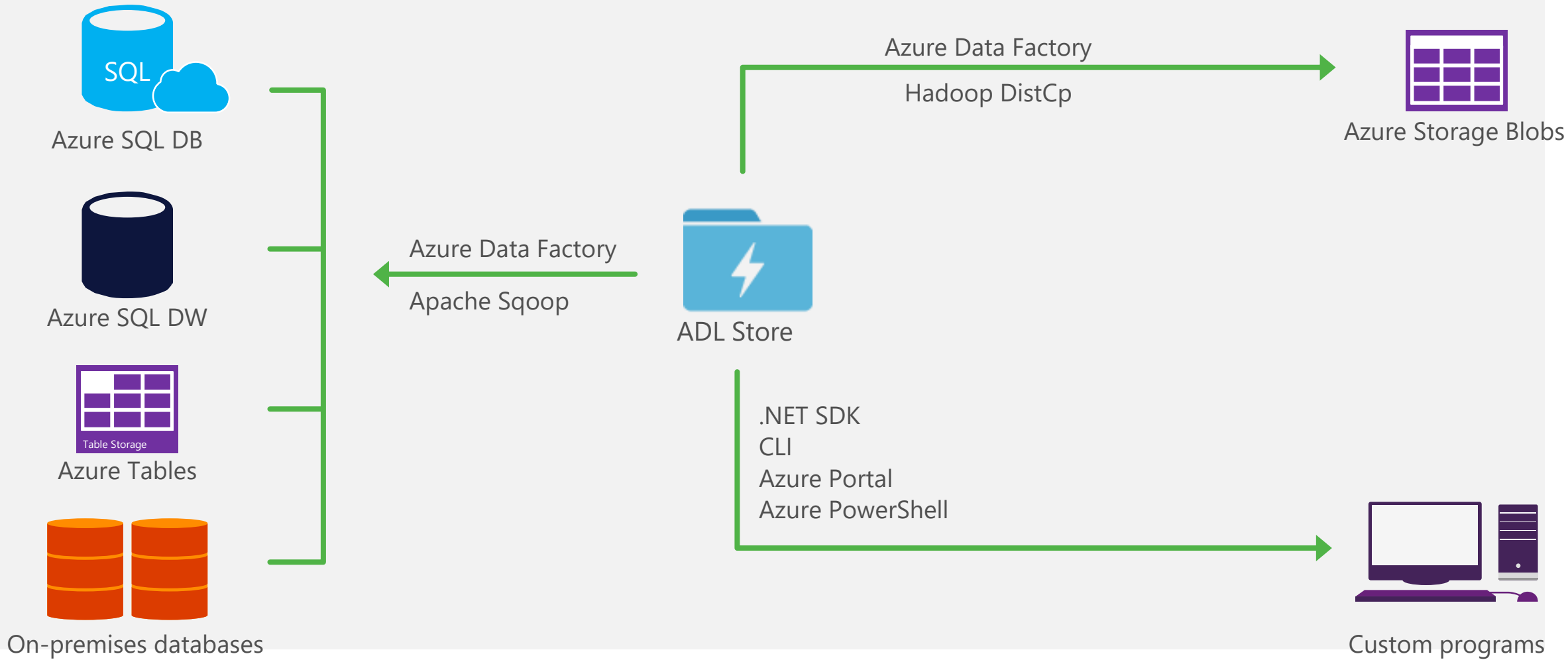
Azure Data Lake Store: Ingress

Data can be ingested into Azure Data Lake Store from a variety of sources

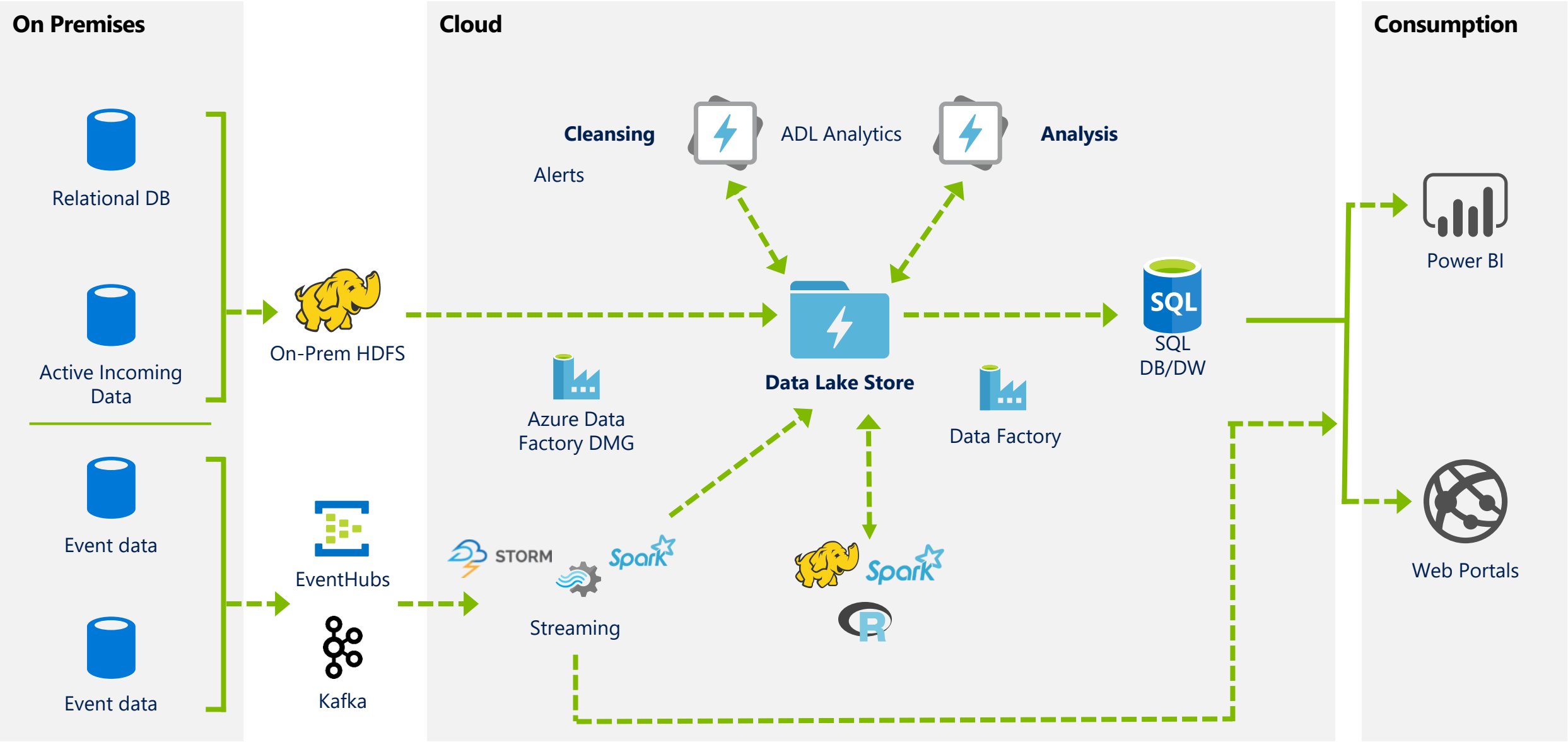


Azure Data Lake Store: Egress

Data can be exported from Azure Data Lake Store into numerous targets/sinks

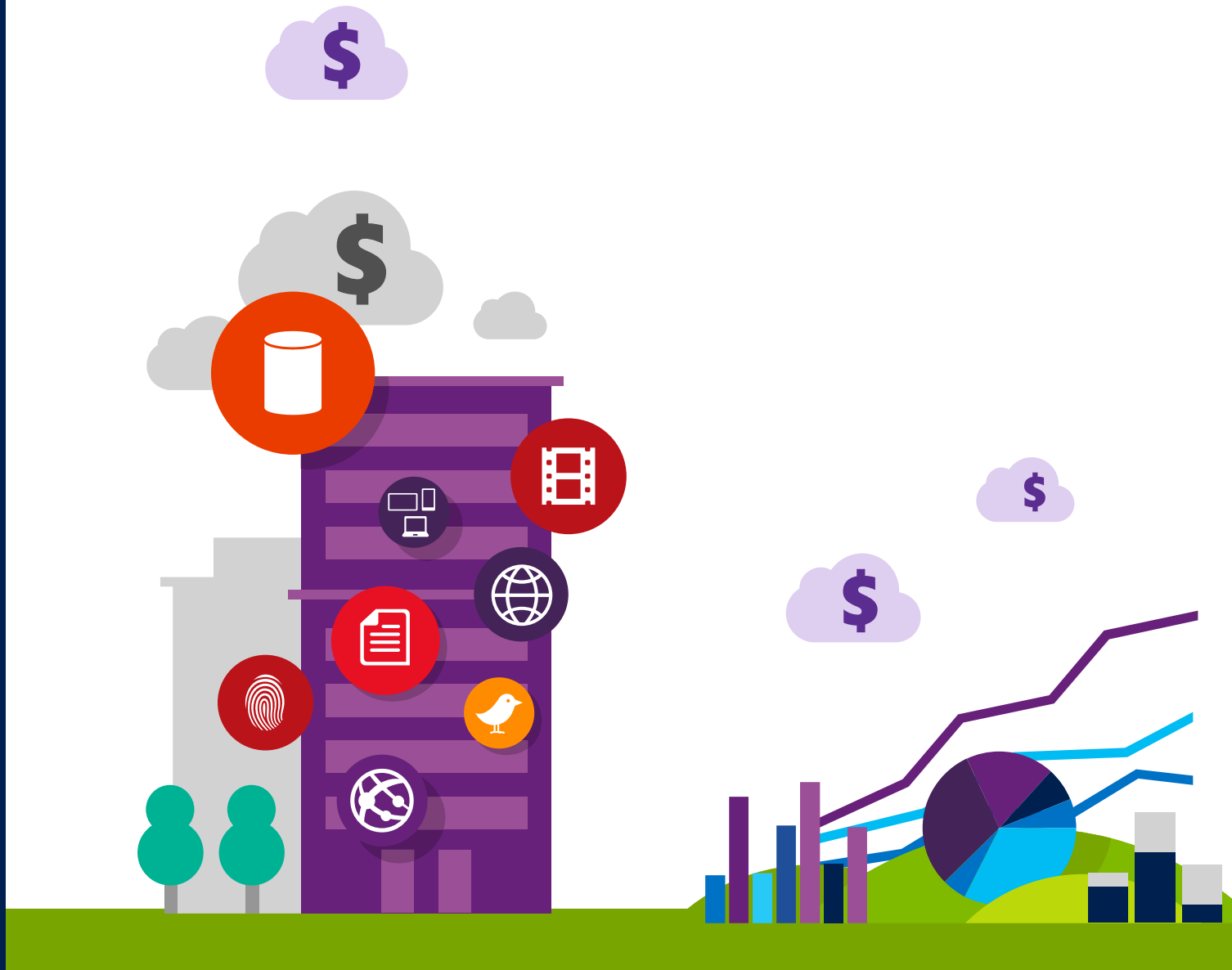


Lambda architecture



Azure Data Lake Store

Costs



Microsoft

Costs breakdown by stage

Ingestion	Number of write transactions
Storage	Data stored per month
Processing	Number of read transactions Number of write transactions
Egress	Number of read transactions

Get all the advantages of ADL Store with cost concepts you are familiar with

Azure Data Lake Analytics

Microsoft

Azure Data Lake Analytics Service

A new distributed analytics service

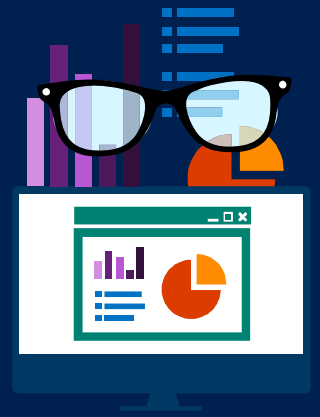
- Built on **Apache YARN**
- **Scales dynamically** with the turn of a dial
- **Pay by the query**
- Supports **Azure AD** for access control, roles, and integration with on-prem identity systems
- Built with **U-SQL** to unify the benefits of SQL with the power of C#
- Processes data **across Azure**

Azure Data Lake Analytics

All data

Productivity from day one

Easy and powerful data preparation

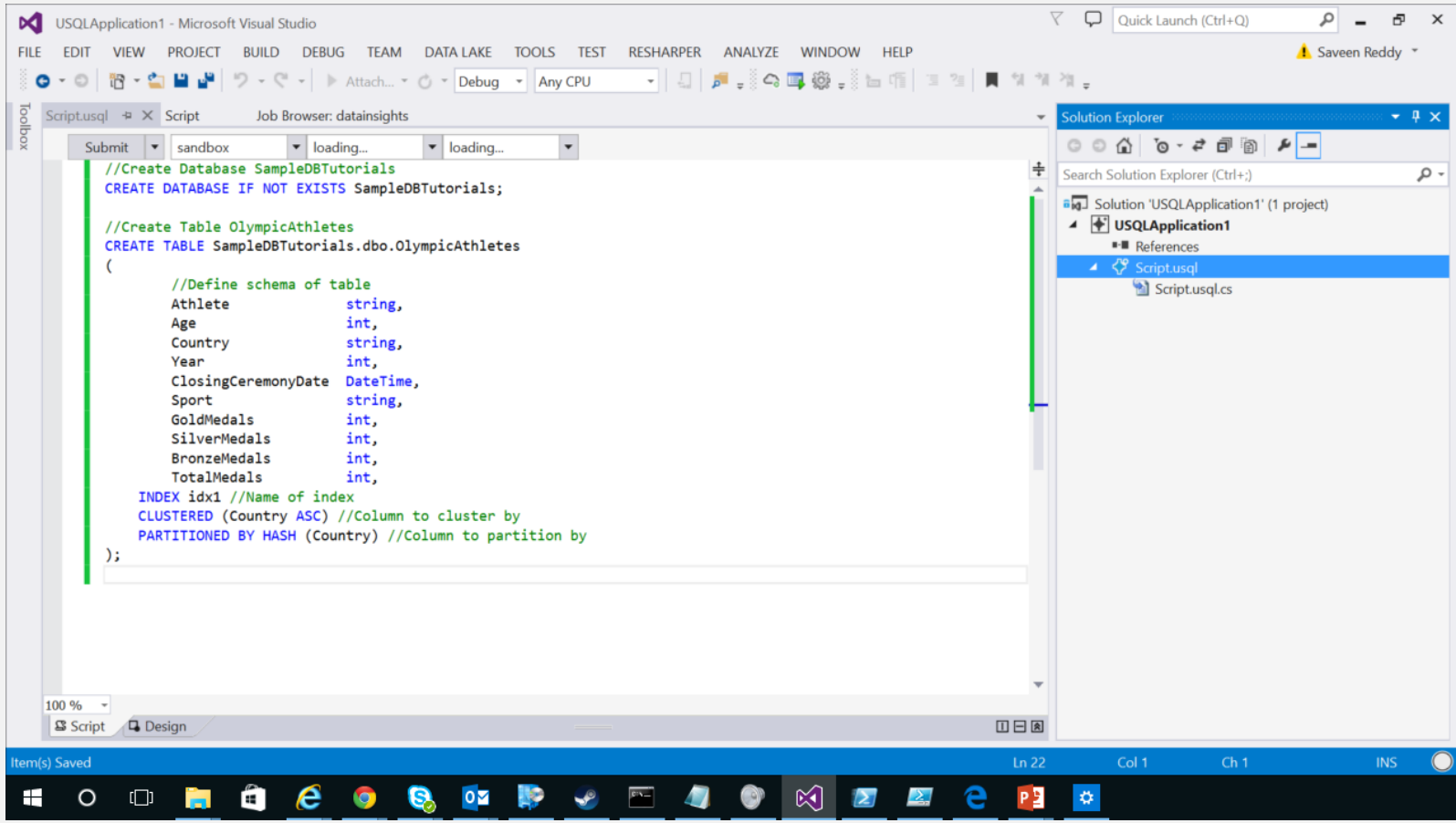


Limitless scale

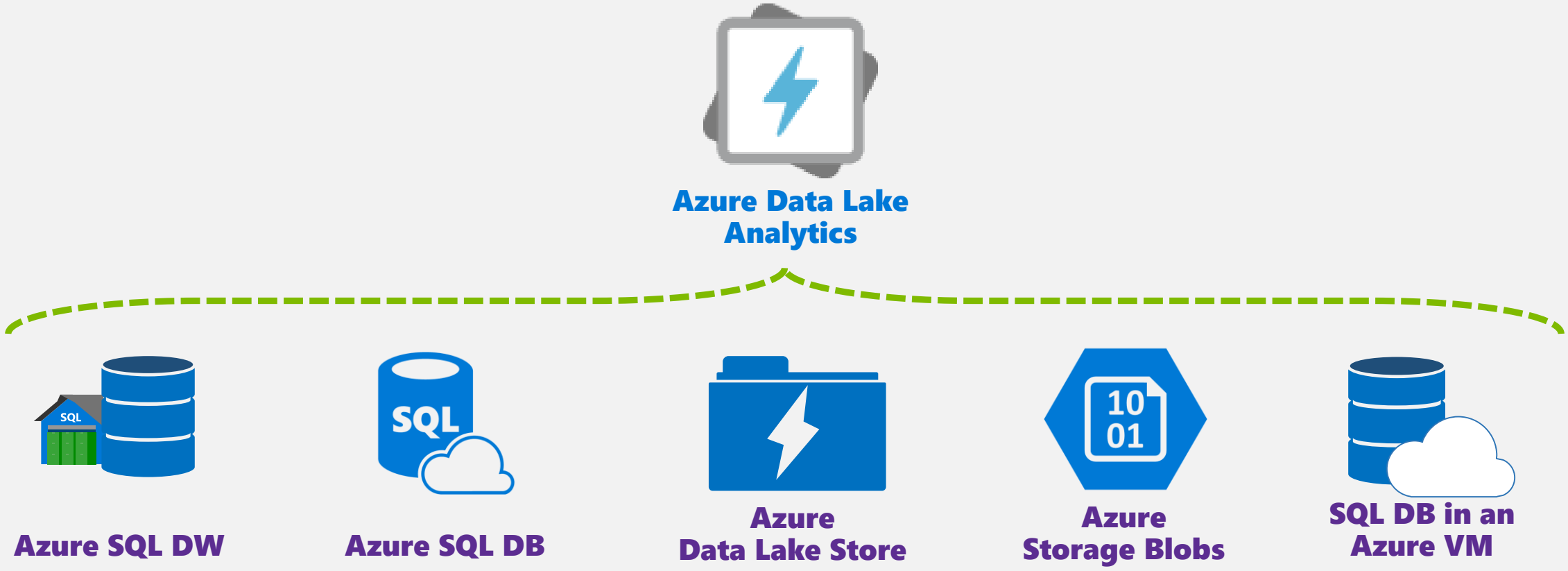
Enterprise-grade

Developing big data apps

- Author, debug, & optimize big data apps in **Visual Studio**
- Multiple Languages **U-SQL, Hive, & Pig**
- Seamlessly integrate **.NET**



Work across all cloud data



Azure Data Lake

U-SQL

Microsoft

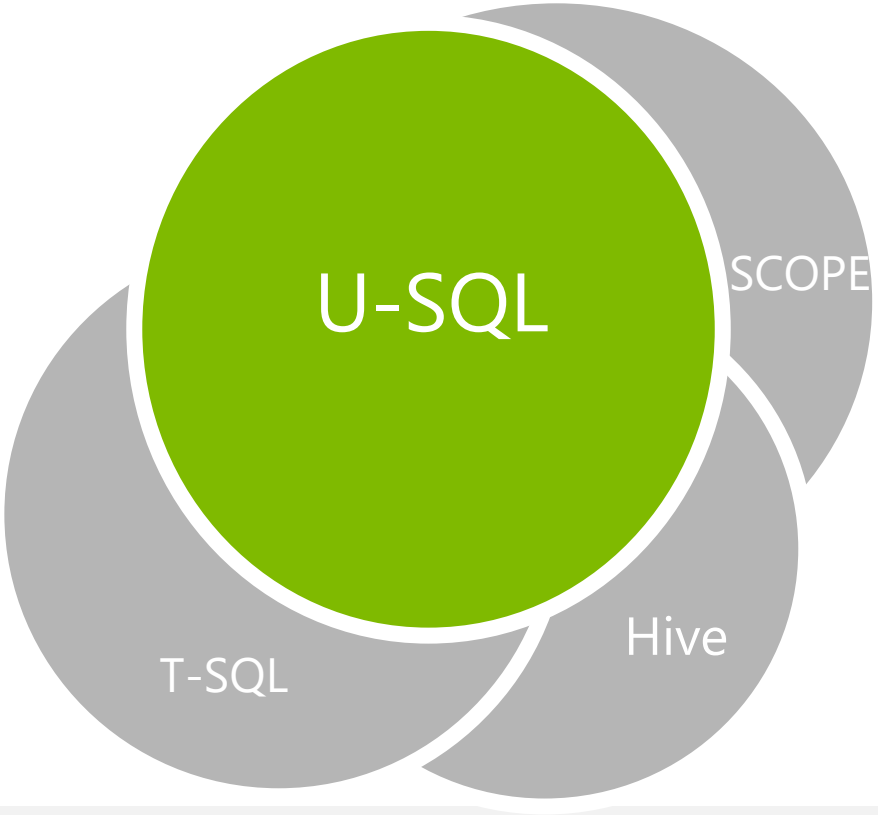
What is U-SQL?

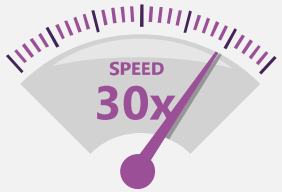
- A **hyper-scalable**, highly extensible language for preparing, transforming and analyzing all data
- Allows users to **focus on the what**—not the how—of business problems
- Built on **familiar languages** (SQL and C#) and supported by a fully integrated development environment
- Built for **data developers & scientists**

The Origins of U-SQL

Next generation large-scale data processing language combining

- The declarative, optimizable and parallelizability of SQL
- The extensibility, expressiveness and familiarity of C#





High performance

Scalable

Affordable

Easy to program

Secure

Usage scenarios

Achieve the same programming experience in batch or interactive

Schematizing unstructured data
(Load-Extract-Transform-Store) for analysis

Cook data for other users (LETS & Share)

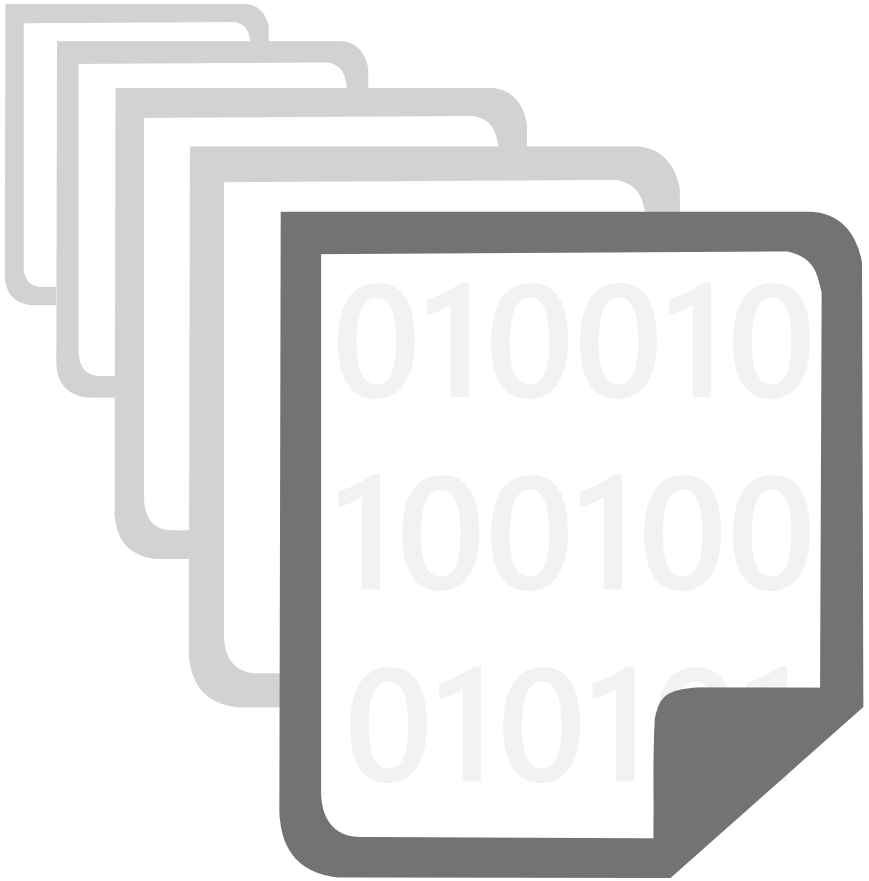
- As unstructured data
- As structured data

Large-scale custom processing with custom code

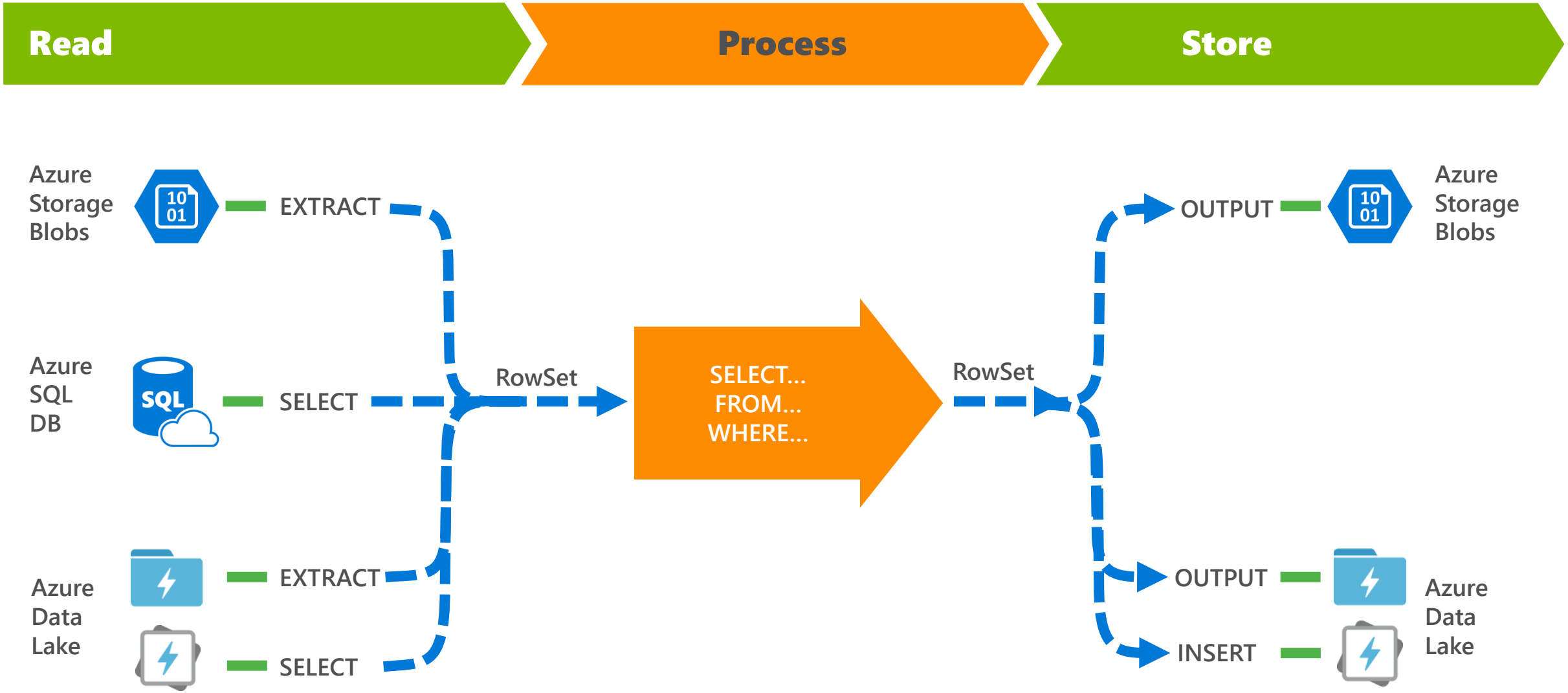
Augment big data with high-value data from where it lives

Expression-flow programming style

- **Automatic** "in-lining" of U-SQL expressions – whole script leads to a single execution model
- Execution plan that is **optimized out-of-the-box** and w/o user intervention
- Per-job and **user-driven** parallelization
- Detail **visibility** into execution steps, for debugging
- **Heat map** functionality to identify performance bottlenecks



U-SQL Queries: General pattern



Anatomy of a U-SQL query

ClassLibrary2 - Microsoft Visual Studio

File Edit View Project Build Debug Team SqlIP Tools Test Analyze Wi

Debug | Any CPU | Start

Server Explorer

```
REFERENCE ASSEMBLY WebLogExtASM;

@rs =
     EXTRACT
       UserID          string,
       Start           DateTime,
       End             DatetTime,
       Region          string,
       SitesVisited    string,
       PagesVisited    string
     FROM "swebhdfs://Logs/WebLogRecords.txt"
     USING WebLogExtractor();

@result = SELECT UserID,
          (End.Subtract(Start)).TotalSeconds AS Duration
          FROM @rs  ORDER BY Duration DESC FETCH 10;

OUTPUT @result TO "swebhdfs://Logs/Results/top10.txt"
USING Outputter.Tsv();
```

Rowset: Conceptually is like an intermediate table… is how U-SQL passes data between statements

- U-SQL types are the same as C# types
- The structure (schema) is first imposed when the data is first extracted/read from the file (schema-on-read)

Input is read from this file in ADL

Custom function to read from input file

C# Expression

Output is stored in this file in ADL

Built-in function that writes the output in TSV format

U-SQL data types

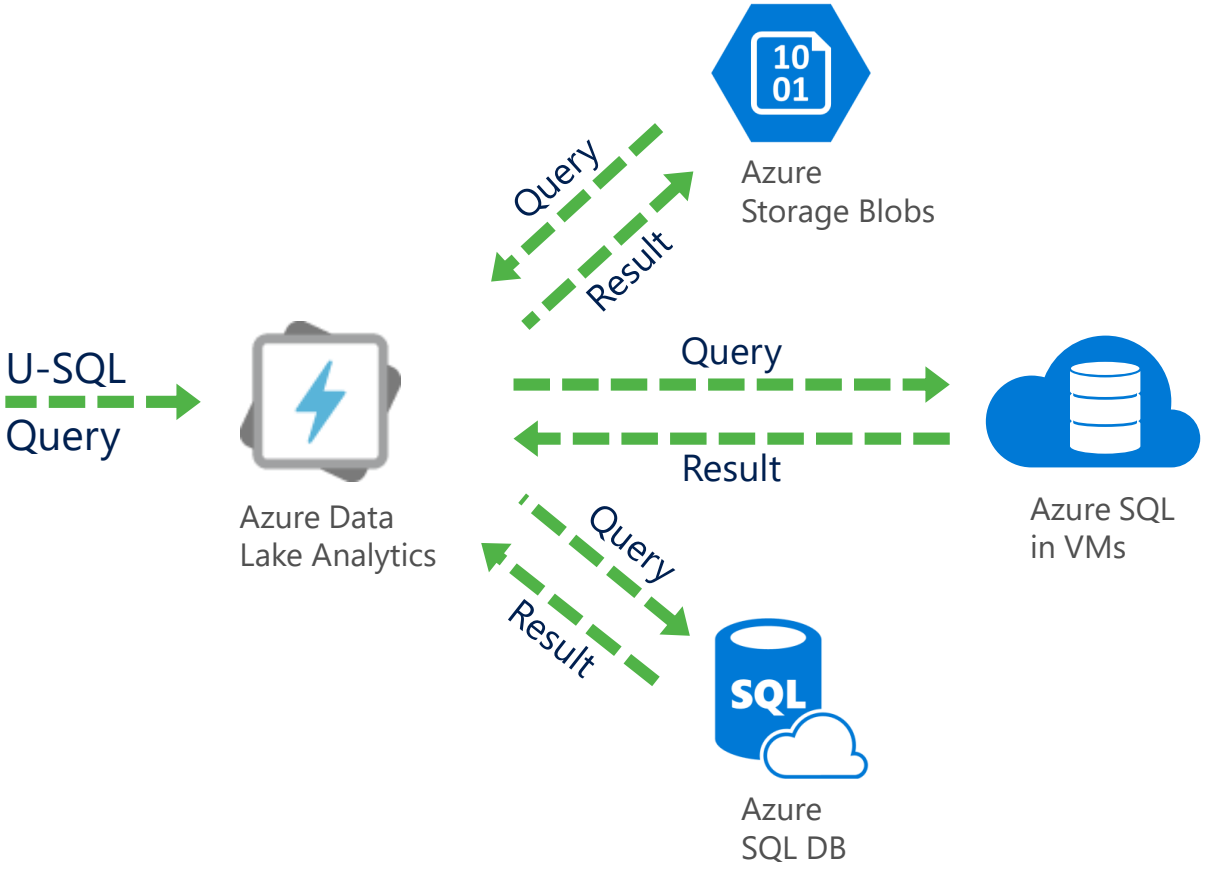
Category	Types	
Numeric	byte, byte? sbyte, sbyte? int, int? uint, unint? long, long? decimal, decimal?	short, short? ushort, ushort? ulong, unlong? float, float? double, double?
Text	char, char? string	
Complex	MAP<K> ARRAY<K,T>	
Temporal	DateTime, DateTime?	
Other	bool, bool? Guid, Guid? Byte[]	

Federated queries: Query data where it lives

Easily query data in multiple Azure data stores without moving it to a single store

Benefits

- Avoid moving large amounts of data across the network between stores
- Single view of data irrespective of physical location
- Minimize data proliferation issues caused by maintaining multiple copies
- Single query language for all data
- Each data store maintains its own sovereignty
- Design choices based on the need



Demo – Lets Run Some Queries

Azure Data Lake Analytics Billing

Microsoft

Azure Data Lake Analytics Billing

- Accounts are **FREE!**
- Pay for the compute resources you want for your **queries**
- Pay for **storage separately**

(query_minutes * parallelism * parallelism_cost_per_minute)+ per_job_charge

Get started today!

For more information visit:
http://azure.com/datalake

Where to learn more…

Microsoft: DAT223.1x Processing Big Data

Course Discussion Progress

Processing Big Data with Azure D

Microsoft Professional Program for Big Data track

10 REQUIRED COURSES | 12-30 HOURS PER COURSE | 8 SKILLS

etail

se runs for three months and starts at the beginning of a quarter. January—March, , July—September, and October —December. The capstone runs for four weeks at g of each quarter: January, April, July, October. For exact dates for the current ease refer to the course detail page on edX.org.

taken during any course run and in any order. When multiple course options ll, only one must be completed to satisfy the requirements for graduation.

Microsoft Virtual Academy | Courses

Advanced | Published: 19 July 2017

Introducing Azure Data Lake

Instructor(s): Saveen Reddy, Nishant Thacker

Micro

Microsoft

Thank you for joining me

Email: GaryHope@Microsoft.com
Twitter: @GaryHope
Mobile: +27 82 7778886

PASS SQLSATURDAY
JOHANNESBURG | 02 SEP 2017