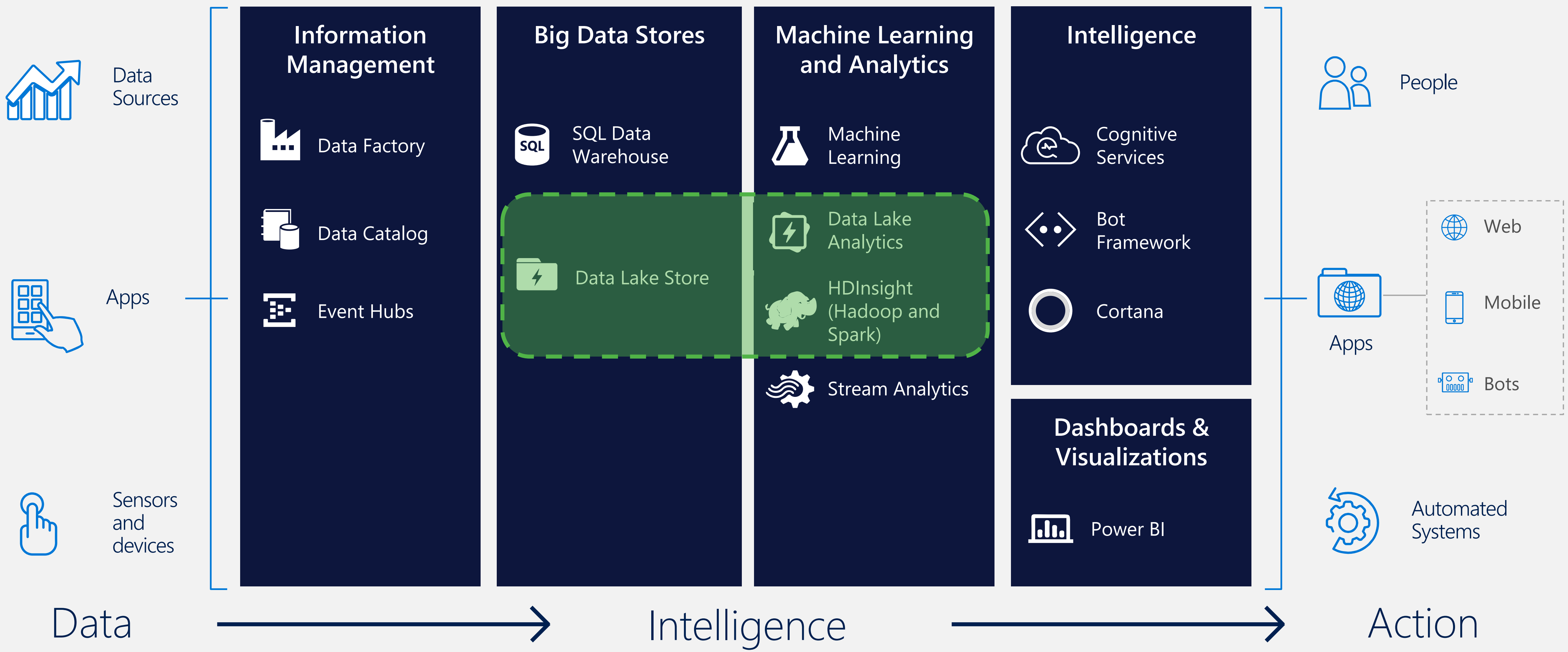
Azure Data Lake Store

Manjunath Suryanarayana

Microsoft

Azure Data Lake

as part of Cortana Intelligence Suite



Data Sources

Apps

Sensors and devices

Information Management

- Data Factory
- Data Catalog
- Event Hubs

Big Data Stores

- SQL Data Warehouse
- Data Lake Store

Machine Learning and Analytics

- Machine Learning
- Data Lake Analytics
- HDInsight (Hadoop and Spark)
- Stream Analytics

Intelligence

- Cognitive Services
- Bot Framework
- Cortana

Dashboards & Visualizations

- Power BI

People

Apps

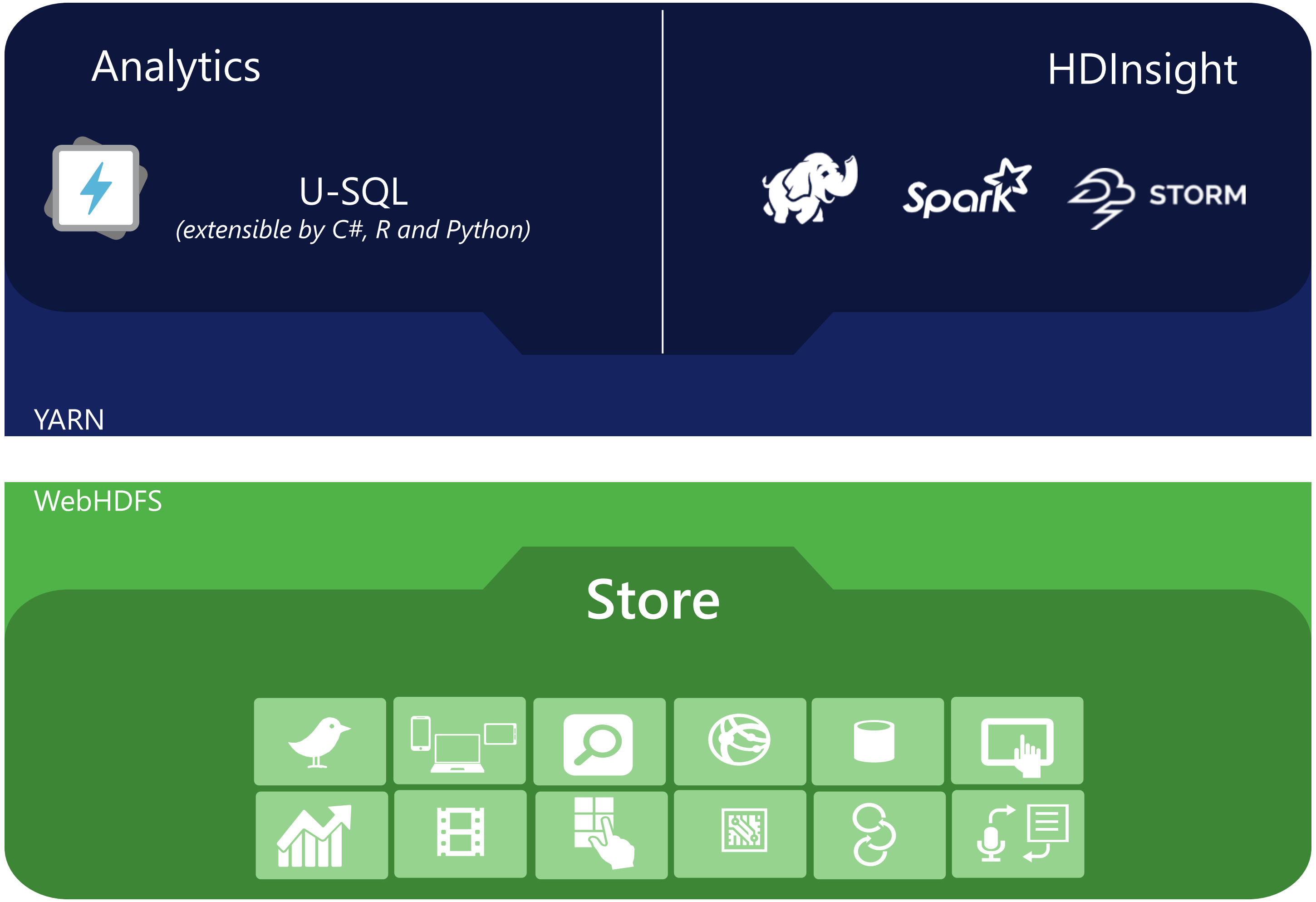
- Web
- Mobile
- Bots

Automated Systems

Data → Intelligence → Action

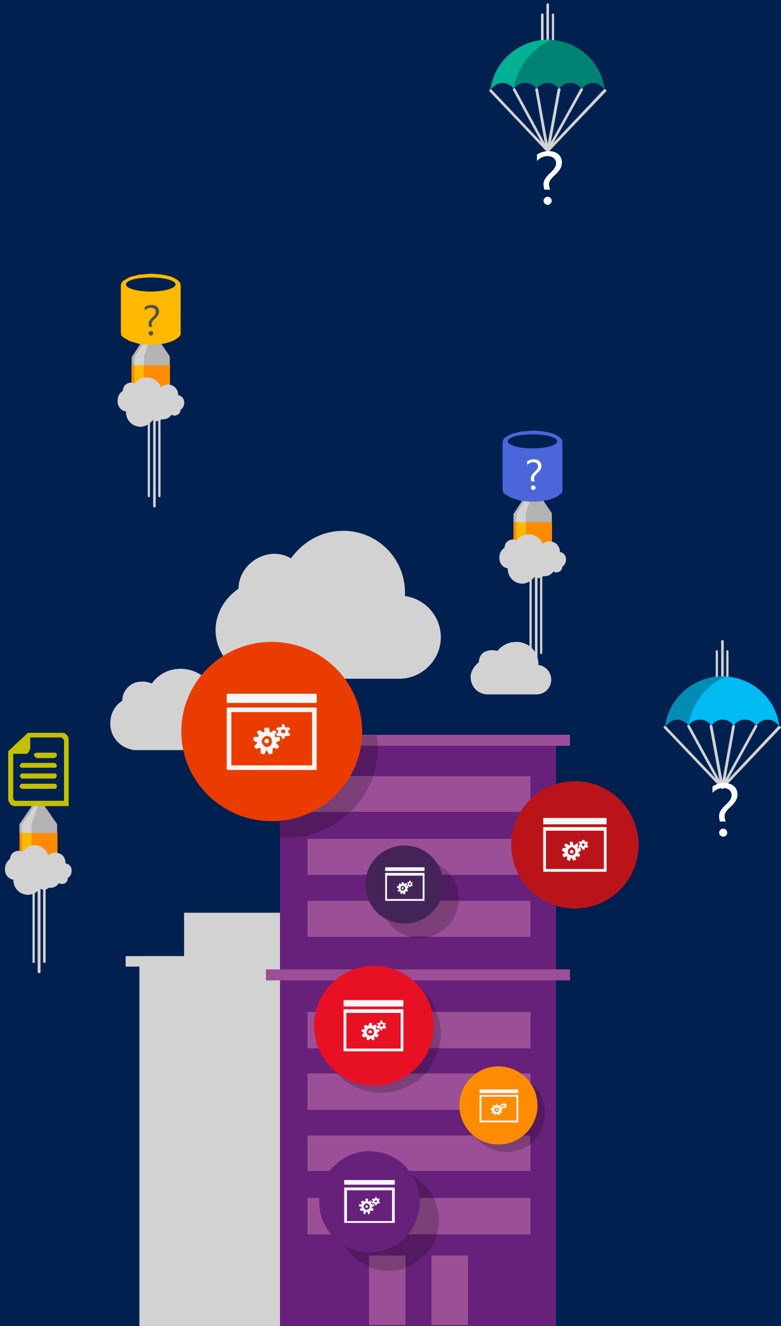
Microsoft

Azure Data Lake



Microsoft

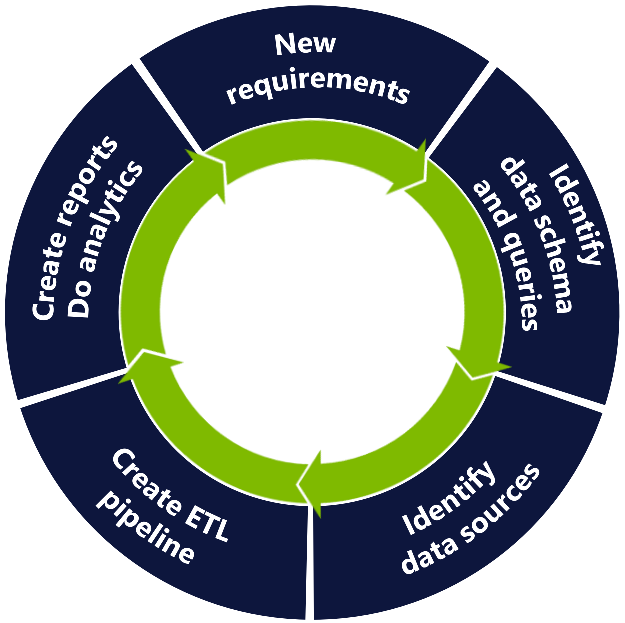
Why data lakes?

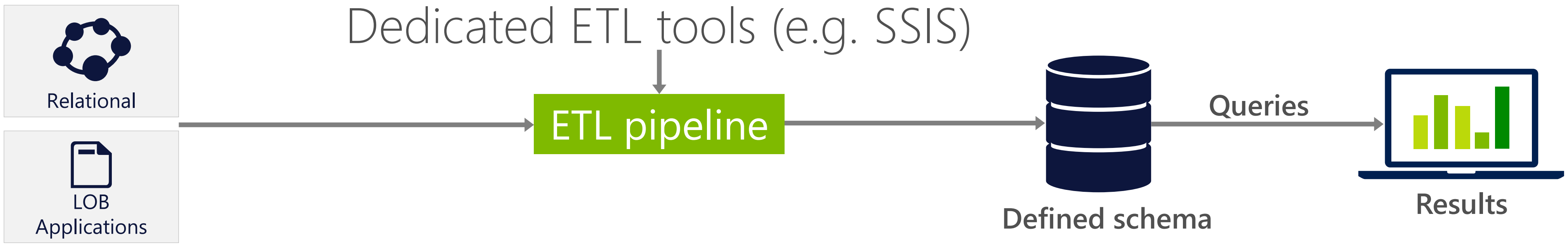


Microsoft

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





Relational

LOB Applications

Dedicated ETL tools (e.g. SSIS)

ETL pipeline

Defined schema

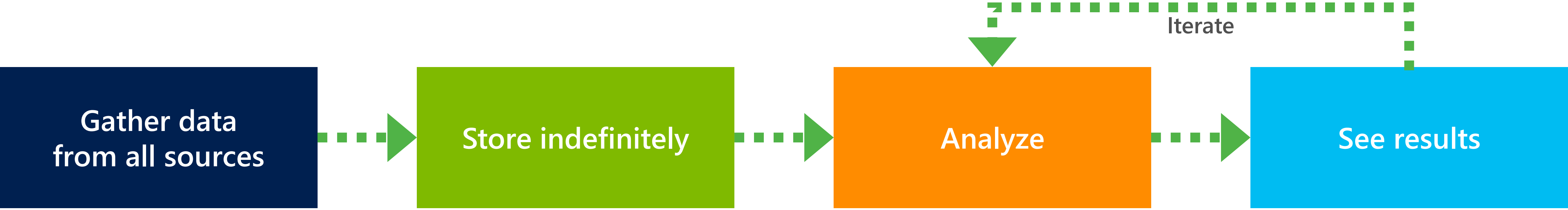
Queries

Results

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



Microsoft

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.

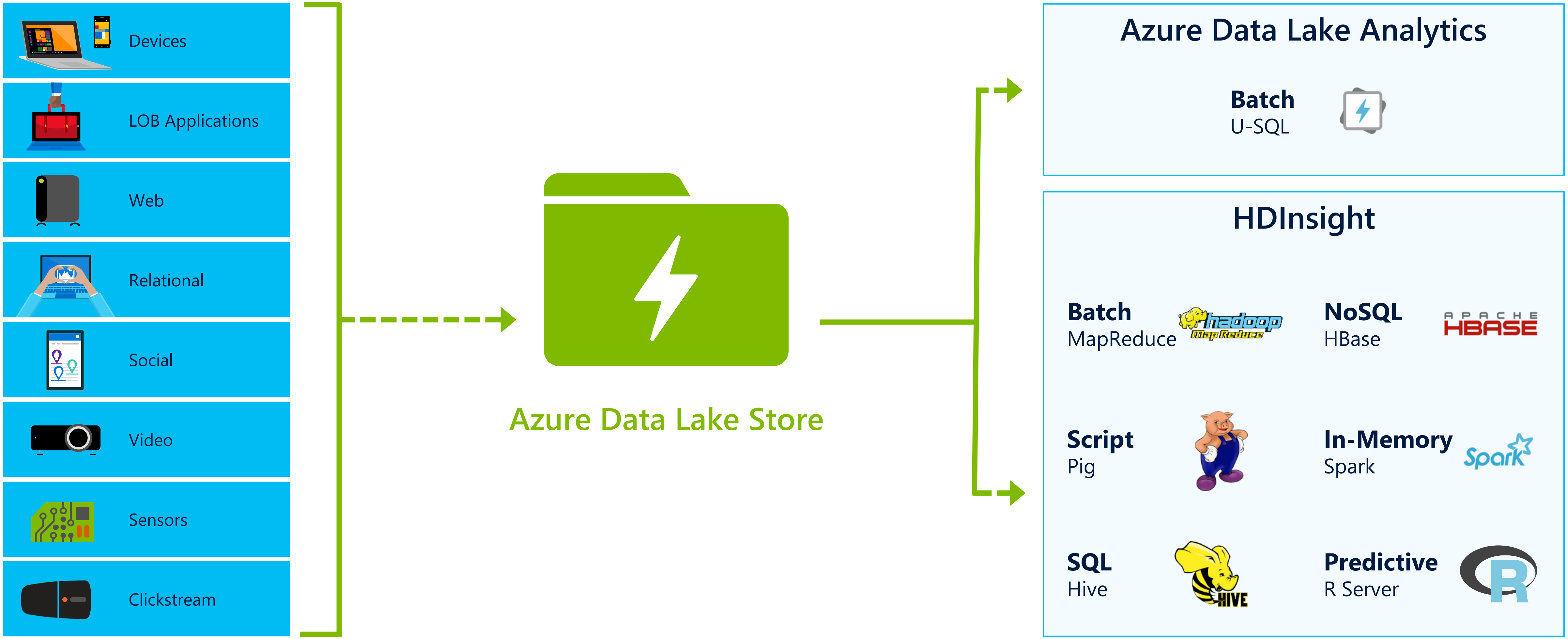
Azure Data Lake Store

Overview

Microsoft

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



Devices

LOB Applications

Web

Relational

Social

Video

Sensors

Clickstream

Azure Data Lake Store

Azure Data Lake Analytics

Batch U-SQL

HDInsight

Batch MapReduce

NoSQL HBase

Script Pig

In-Memory Spark

SQL Hive

Predictive R Server

Scale, performance, reliability

Microsoft

Azure Data Lake Store: no scale limits

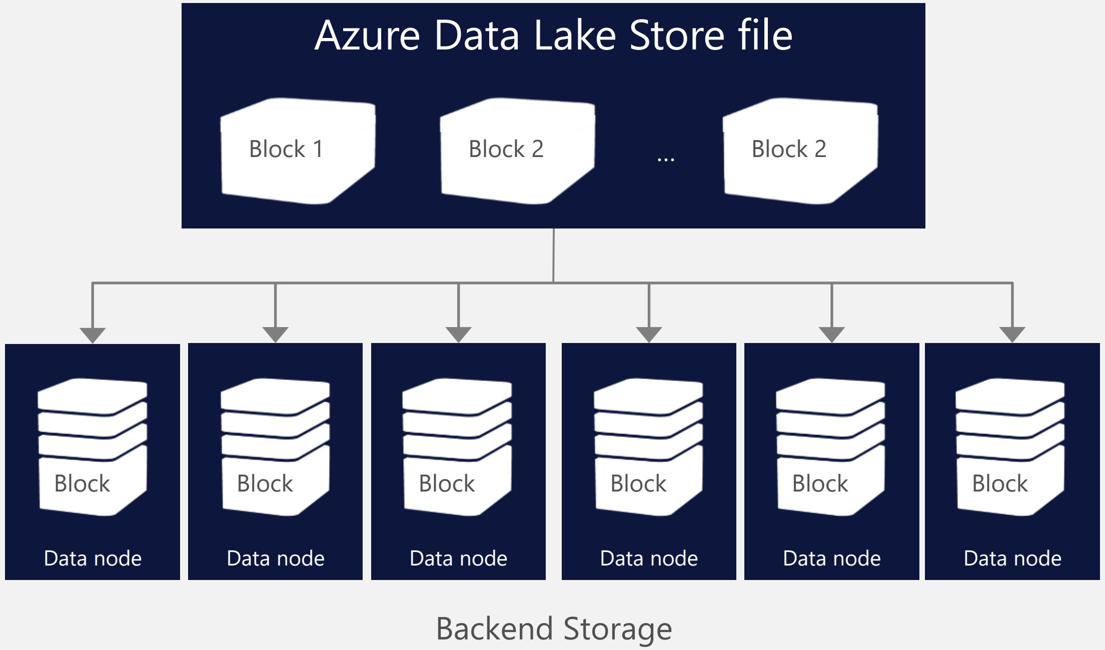
Azure Data Lake Store integrates with Azure Active Directory (AAD) for:

- 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

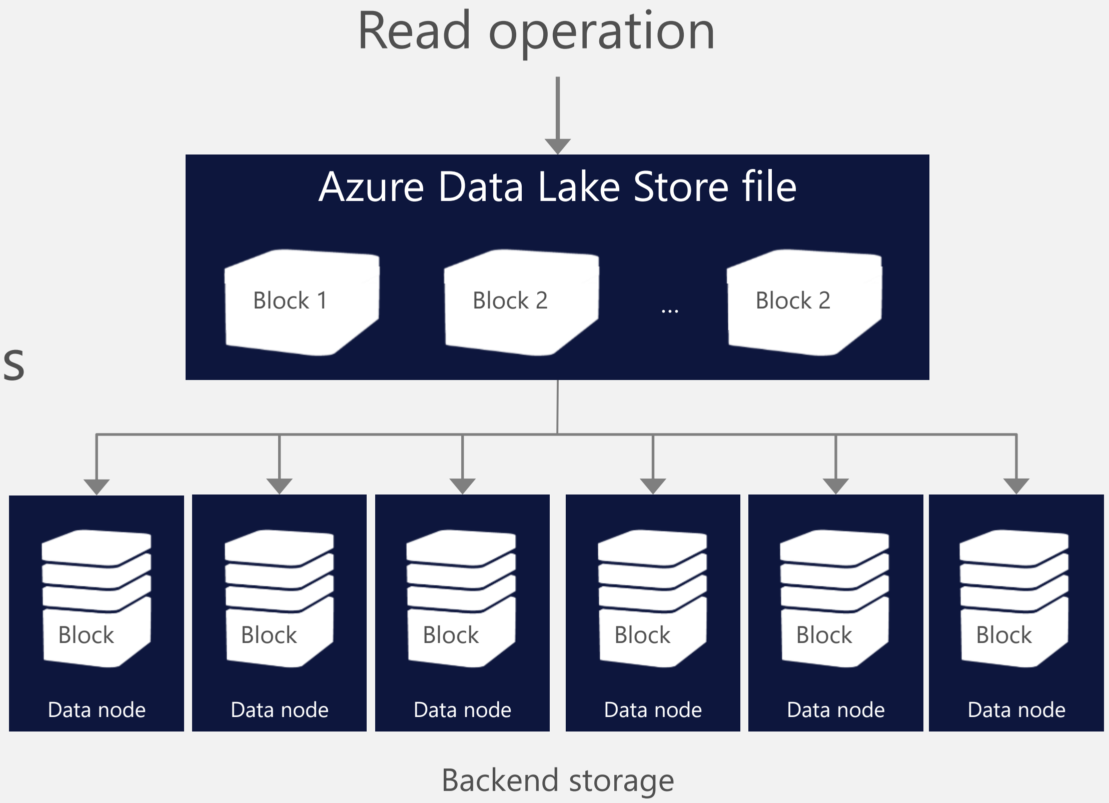
ADL Store Unlimited Scale – How it works

- 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.



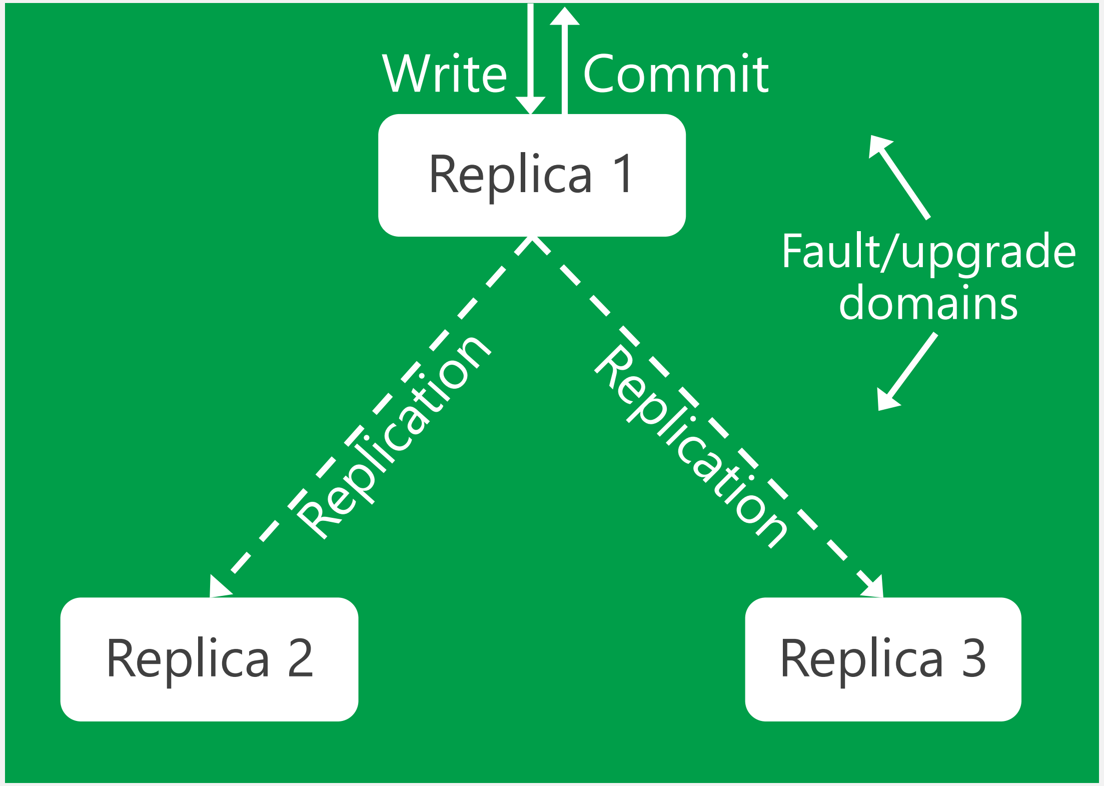
ADL Store offers massive throughput

- 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



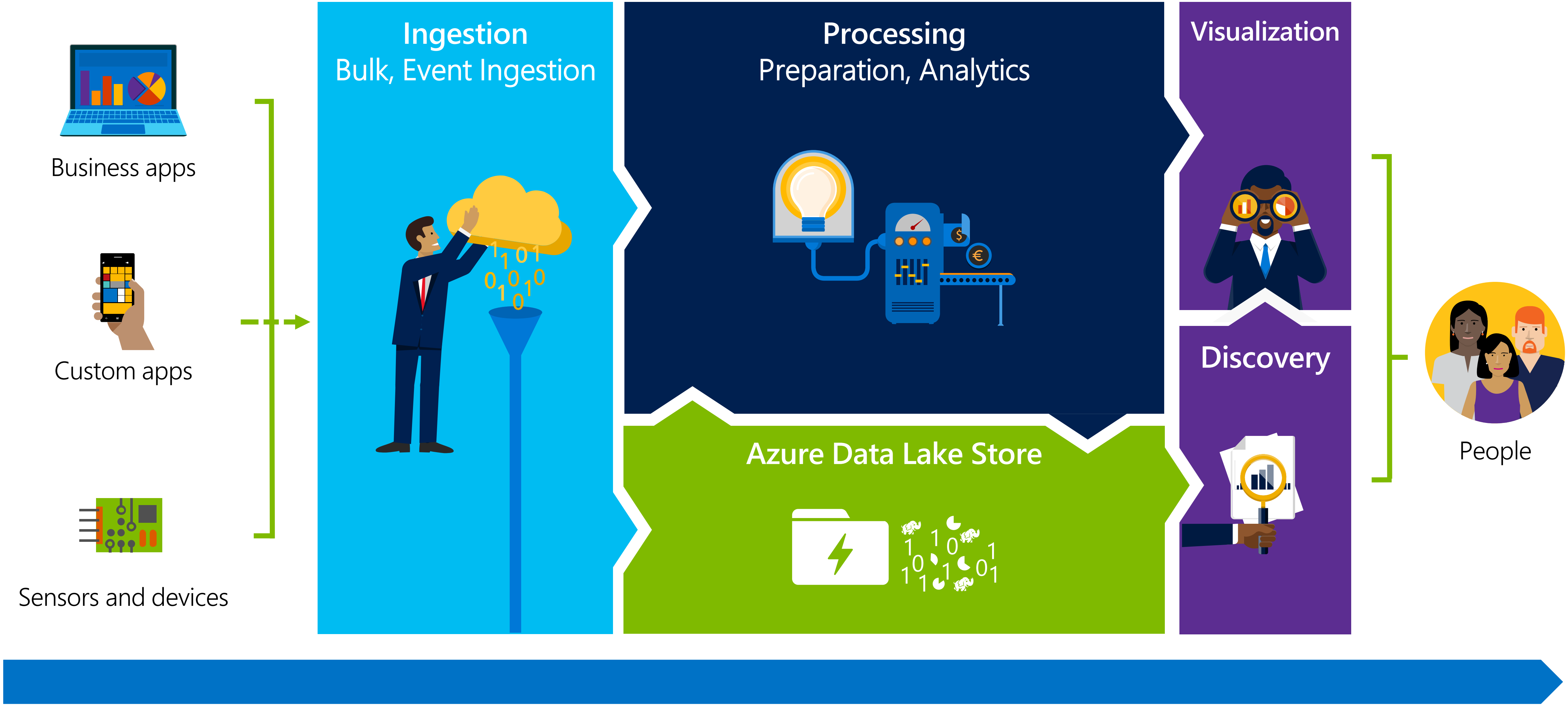
Data is never lost or unavailable even under failures

The building blocks

Ingestion, processing, egress, visualization, and orchestration tools

Microsoft

Big Data Flow



Business apps

Custom apps

Sensors and devices

Ingestion
Bulk, Event Ingestion

Processing
Preparation, Analytics

Azure Data Lake Store

Visualization

Discovery

People

Ingestion tools – Getting started

Data on your desktop

Azure Portal

Easy to use

Good for small amount of data

Analyzing data using Portal

PowerShell

Upload file and folders

Control parallelism

Control format of upload

Need to use other services

ADL Tools for Visual Studio

Integrated experience

Drag-and-drop

Programmatic Analytics

CLI

Linux, Mac

Most features of PowerShell

Data located in other stores

Azure Data Factory

Copy Wizard for intuitive one-time copy from multiple sources

AdlCopy

Copy data easily from Azure Storage at least cost

OSS tools on HDI

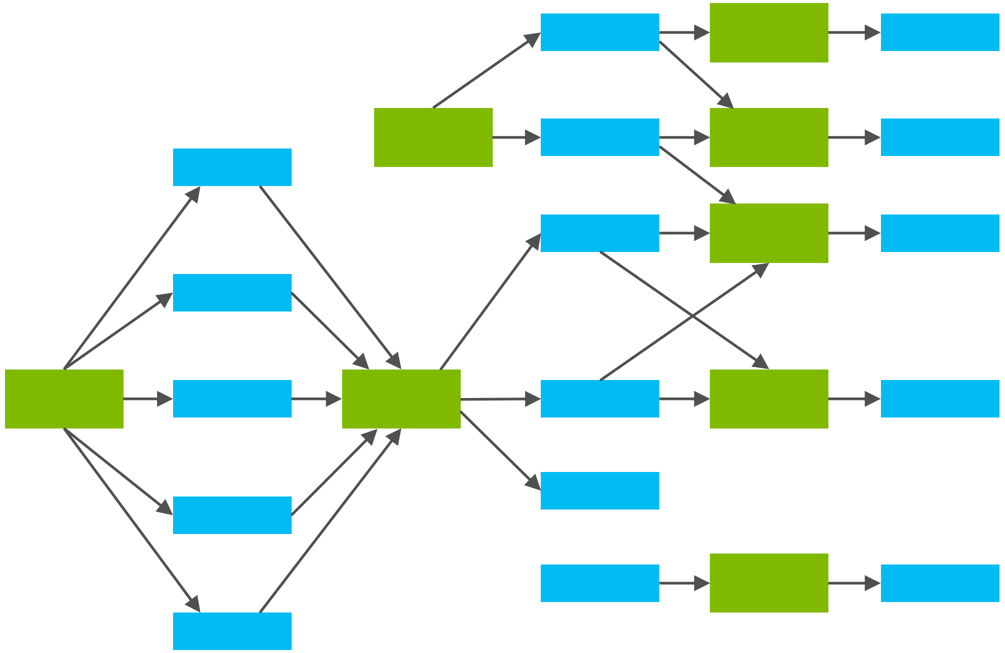
Distcp, Sqoop

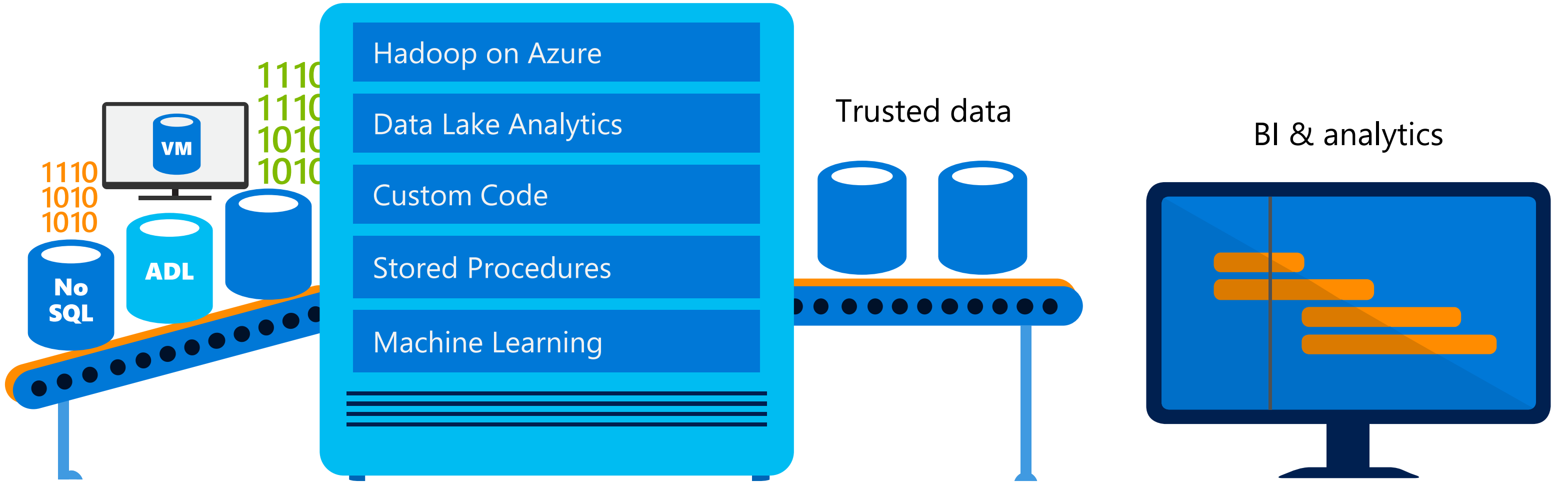
If analyzing data using HDInsight

Azure Data Factory

Compose, orchestrate & monitor data services at scale

- Fully managed service
- Any data on-premises or in the cloud
- Single pane of glass management
- Global service infrastructure
- Cost Effective





Hadoop on Azure

Data Lake Analytics

Custom Code

Stored Procedures

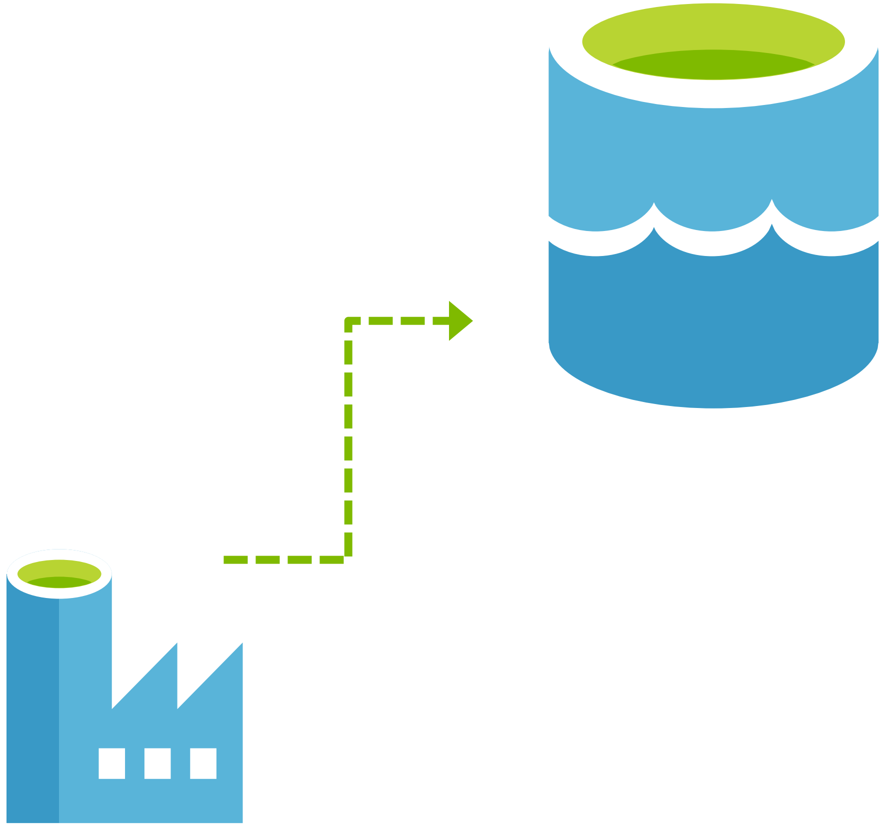
Machine Learning

Trusted data

BI & analytics

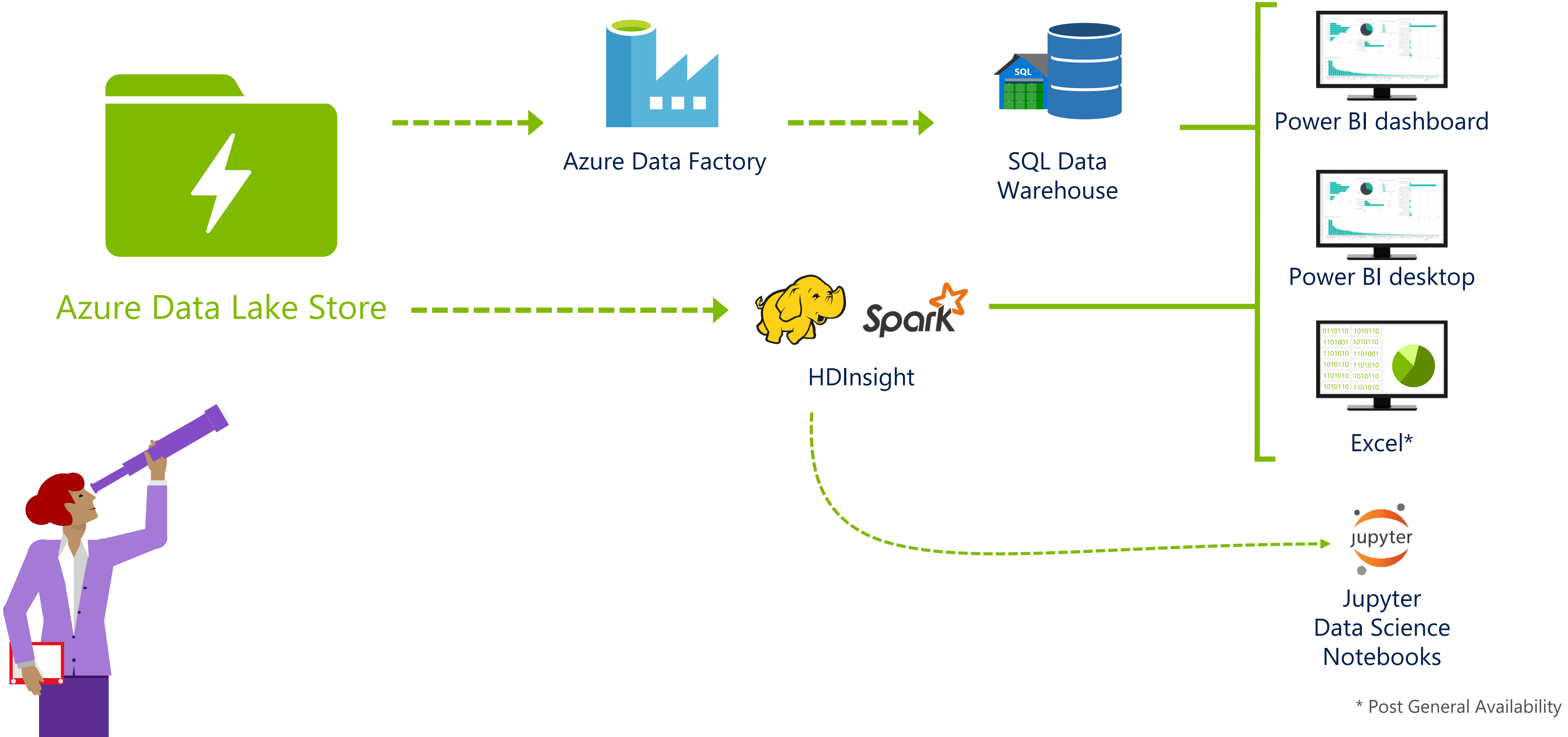
Azure Data Factory

Connects ADL Store out-of-the-box to all your stores



Category	Data store	Supported as source	Supported as sink
Azure	**Azure Data Lake Store**	●	●
	Azure Blob storage	●	●
	Azure SQL Database	●	●
	Azure SQL Data Warehouse	●	●
	Azure Table storage	●	●
	Azure DocumentDB	●	●
Databases	SQL Server*	●	●
	Oracle*	●	●
	MySQL*	●	
	DB2*	●	
	Teradata*	●	
File	HDFS*	●	
	Others	●	

Visualizing data



Azure Data Lake Store

Azure Data Factory

SQL Data Warehouse

HDInsight

Power BI dashboard

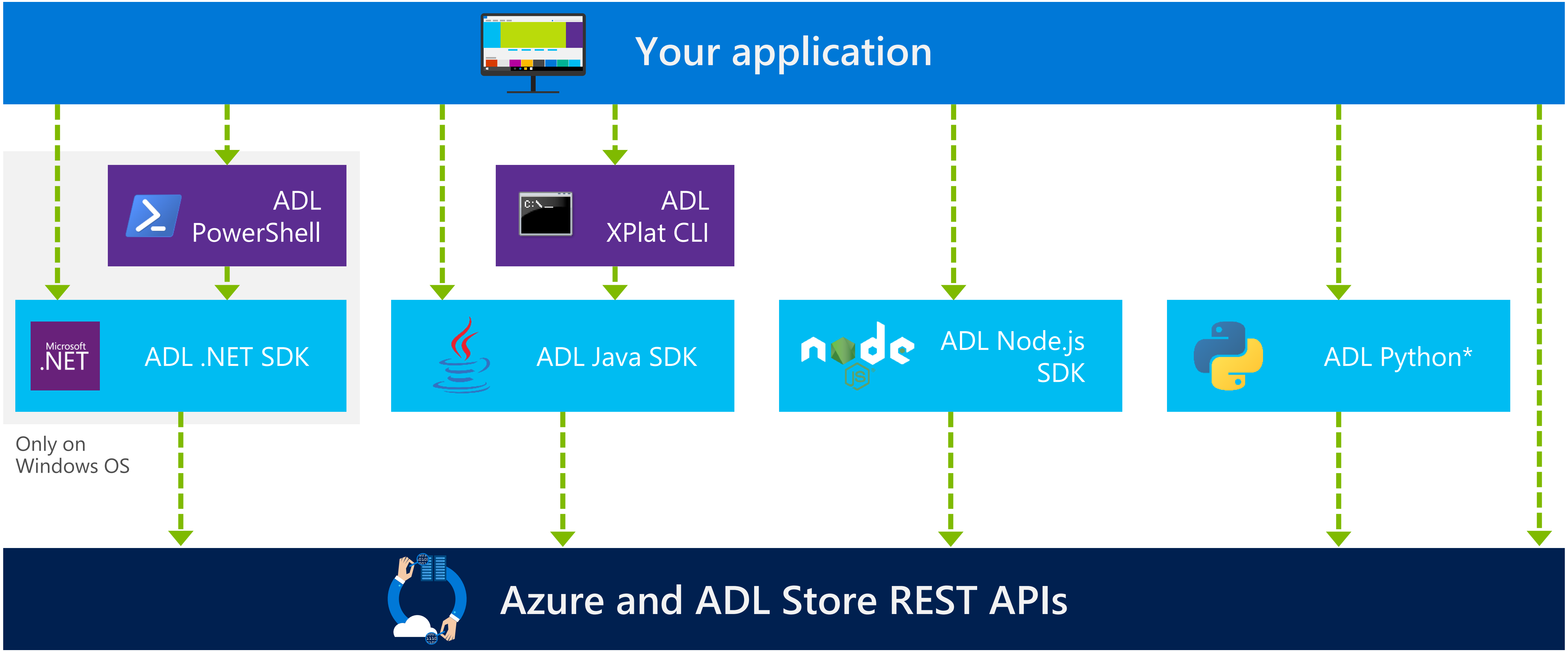
Power BI desktop

Excel*

Jupyter Data Science Notebooks

* Post General Availability

Customizing using SDKs/APIs



Your application

ADL PowerShell

ADL XPlat CLI

ADL .NET SDK

ADL Java SDK

ADL Node.js SDK

ADL Python*

Only on Windows OS

Azure and ADL Store REST APIs

* At General Availability

Building pipelines - Management and orchestration

Out-of-the-box tools

Azure Data Factory

First-class support

Azure Stream Analytics

Seamlessly stream data

OSS tools

Supports OSS tools

PowerShell

Management with Workflow & Script Runbooks

Custom tools

ADL Store SDK

Available in multiple languages

REST APIs

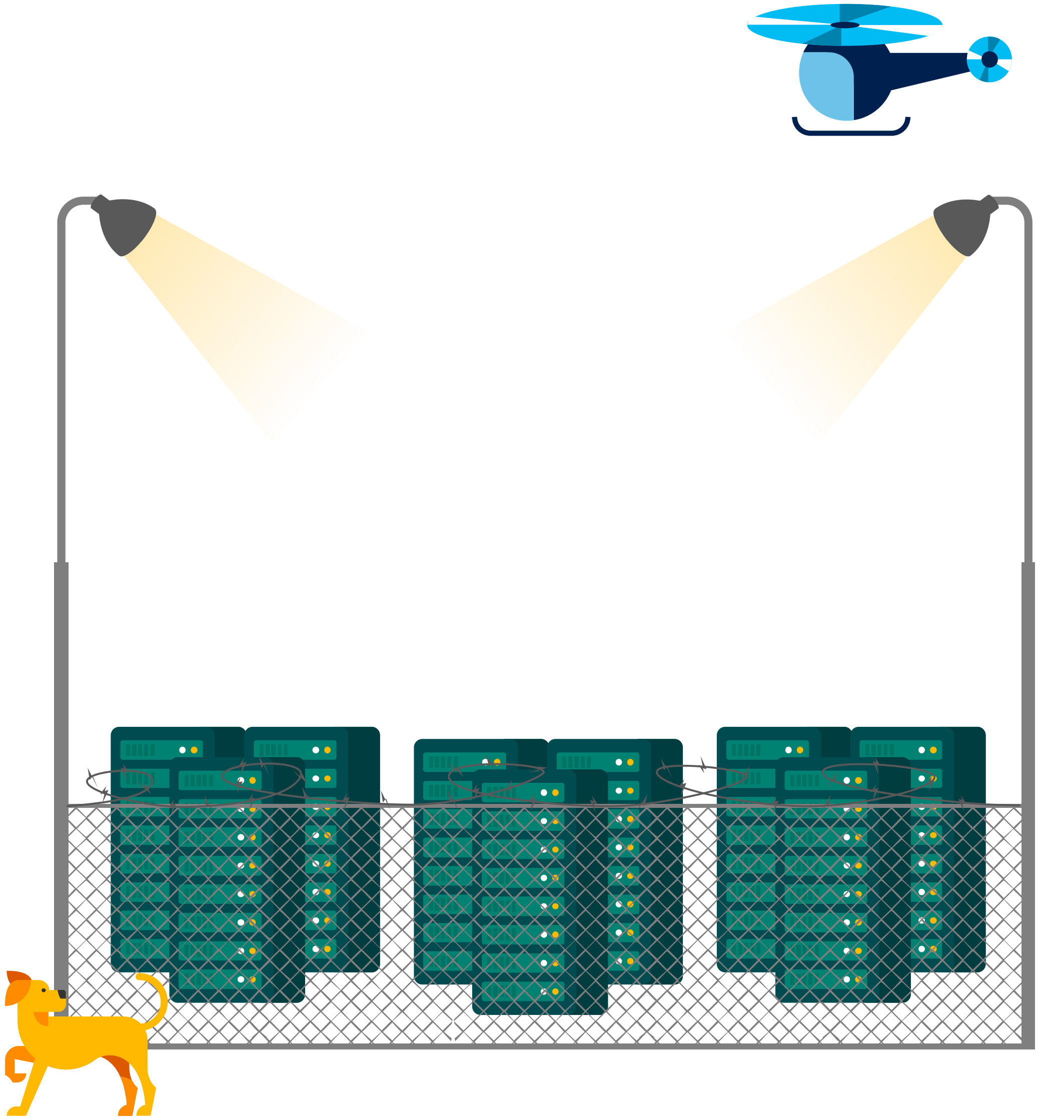
For unsupported languages and platforms

Security

Microsoft

Security features

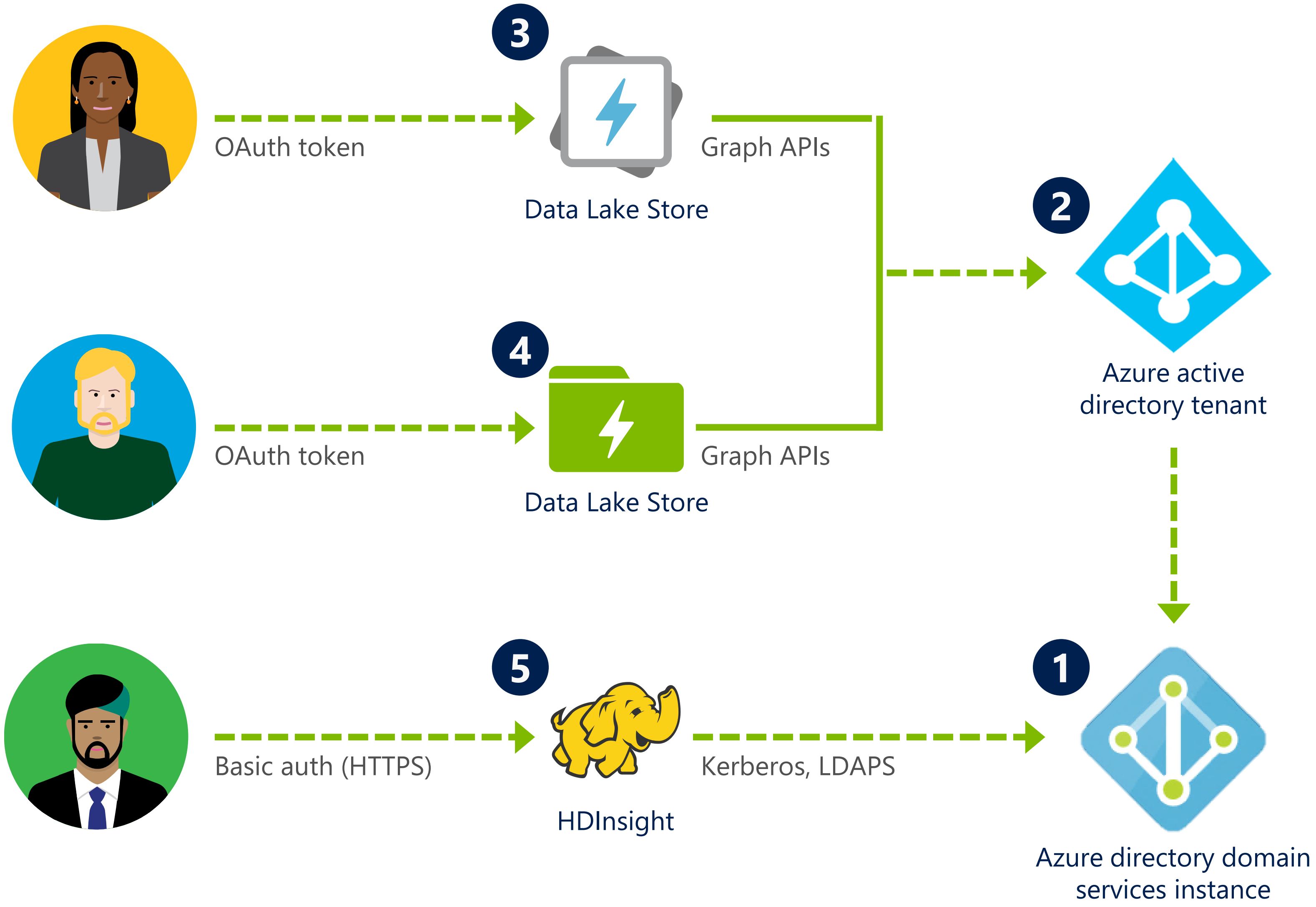
Identity Management & Authentication	Azure Active Directory
Access Control & Authorization	Azure RBAC for Account Management File & Folder level POSIX ACLs
Auditing	Azure Diagnostic Audit Logs
Data Protection & Encryption	Encryption on the wire using HTTPS Transparent Service side encryption using service & customer managed keys



ADL 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

Leveraging Azure Active Directory



1 Create ADDS instance in separate VNET

2 Add users to AAD Tenant

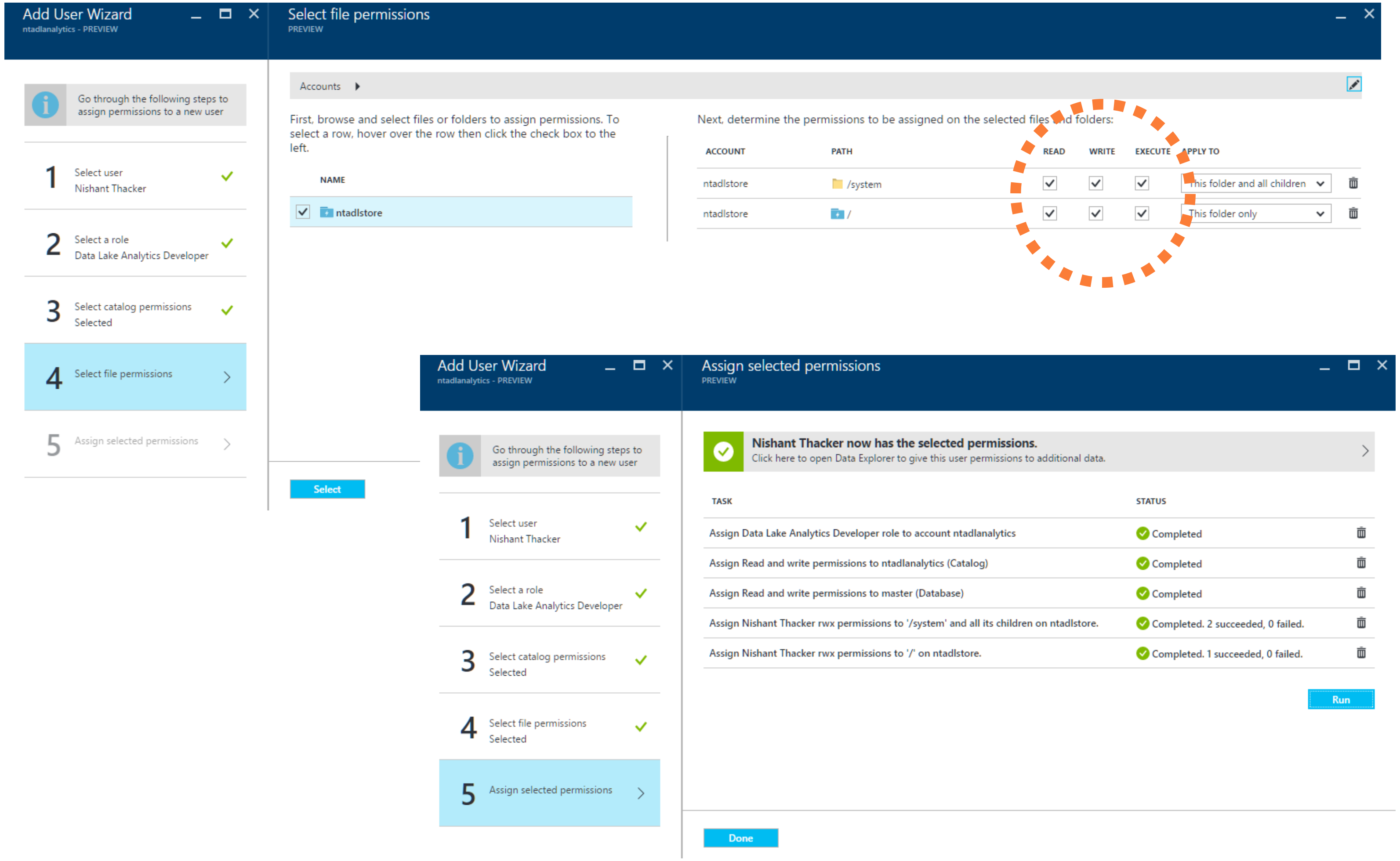
3 Add users to ADLA RBAC roles

4 Add users to ADLS RBAC roles & file system ACLs

5 Join HDInsight cluster to ADDS instance

ADL 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



Add User Wizard
ntadlanalytics - PREVIEW

Go through the following steps to assign permissions to a new user

1 Select user
Nishant Thacker ✓

2 Select a role
Data Lake Analytics Developer ✓

3 Select catalog permissions
Selected ✓

4 Select file permissions

5 Assign selected permissions

Select file permissions
PREVIEW

Accounts

First, browse and select files or folders to assign permissions. To select a row, hover over the row then click the check box to the left.

NAME
☑ ntadlstore

Next, determine the permissions to be assigned on the selected files and folders:

ACCOUNT	PATH	READ	WRITE	EXECUTE	APPLY TO
ntadlstore	/system	☑	☑	☑	This folder and all children
ntadlstore	/	☑	☑	☑	This folder only

Add User Wizard
ntadlanalytics - PREVIEW

Go through the following steps to assign permissions to a new user

1 Select user
Nishant Thacker ✓

2 Select a role
Data Lake Analytics Developer ✓

3 Select catalog permissions
Selected ✓

4 Select file permissions
Selected ✓

5 Assign selected permissions

Assign selected permissions
PREVIEW

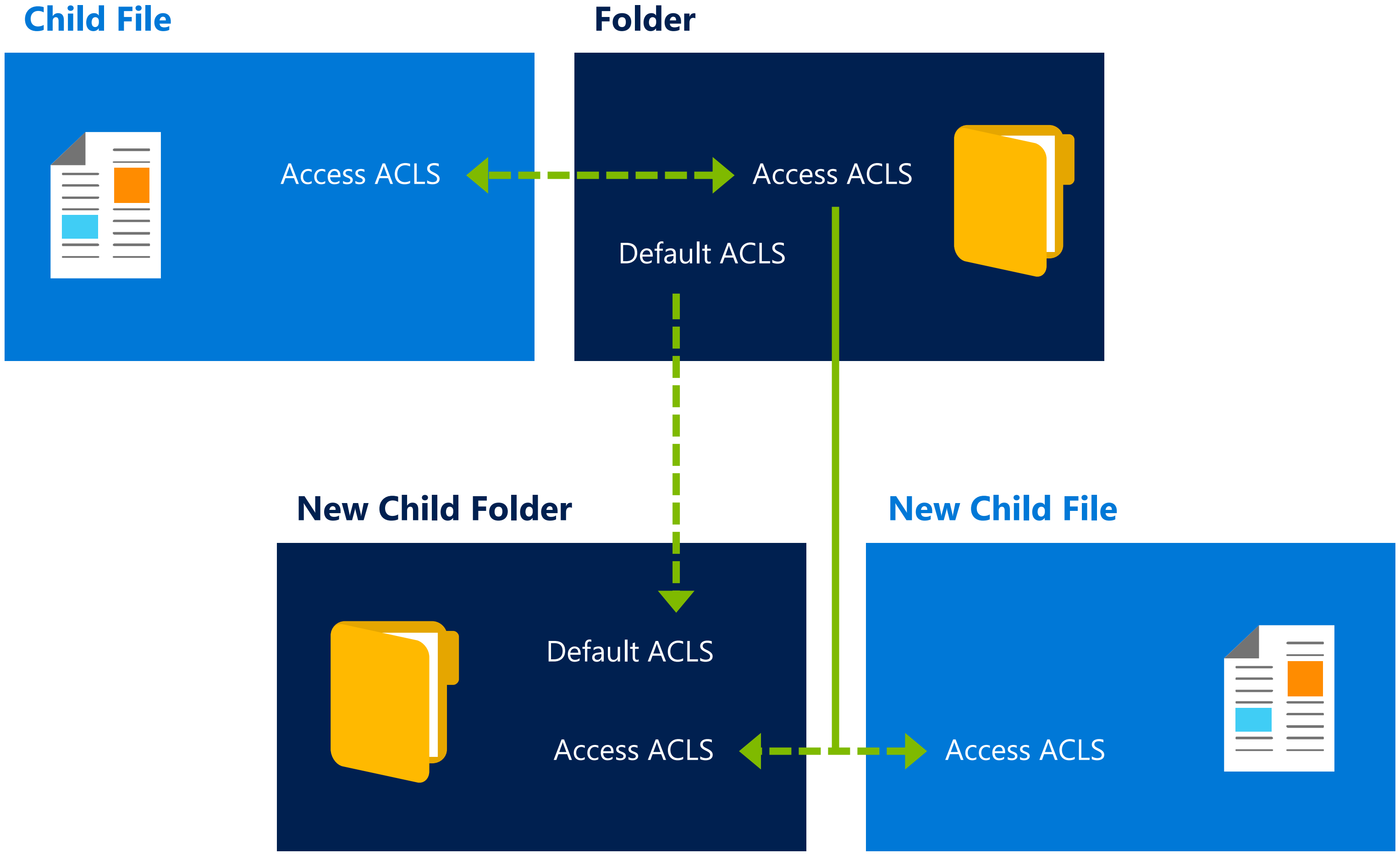
Nishant Thacker now has the selected permissions.
Click here to open Data Explorer to give this user permissions to additional data.

TASK	STATUS
Assign Data Lake Analytics Developer role to account ntadlanalytics	✓ Completed
Assign Read and write permissions to ntadlanalytics (Catalog)	✓ Completed
Assign Read and write permissions to master (Database)	✓ Completed
Assign Nishant Thacker rwx permissions to '/system' and all its children on ntadlstore.	✓ Completed. 2 succeeded, 0 failed.
Assign Nishant Thacker rwx permissions to '/' on ntadlstore.	✓ Completed. 1 succeeded, 0 failed.

Granular control of file and folder access

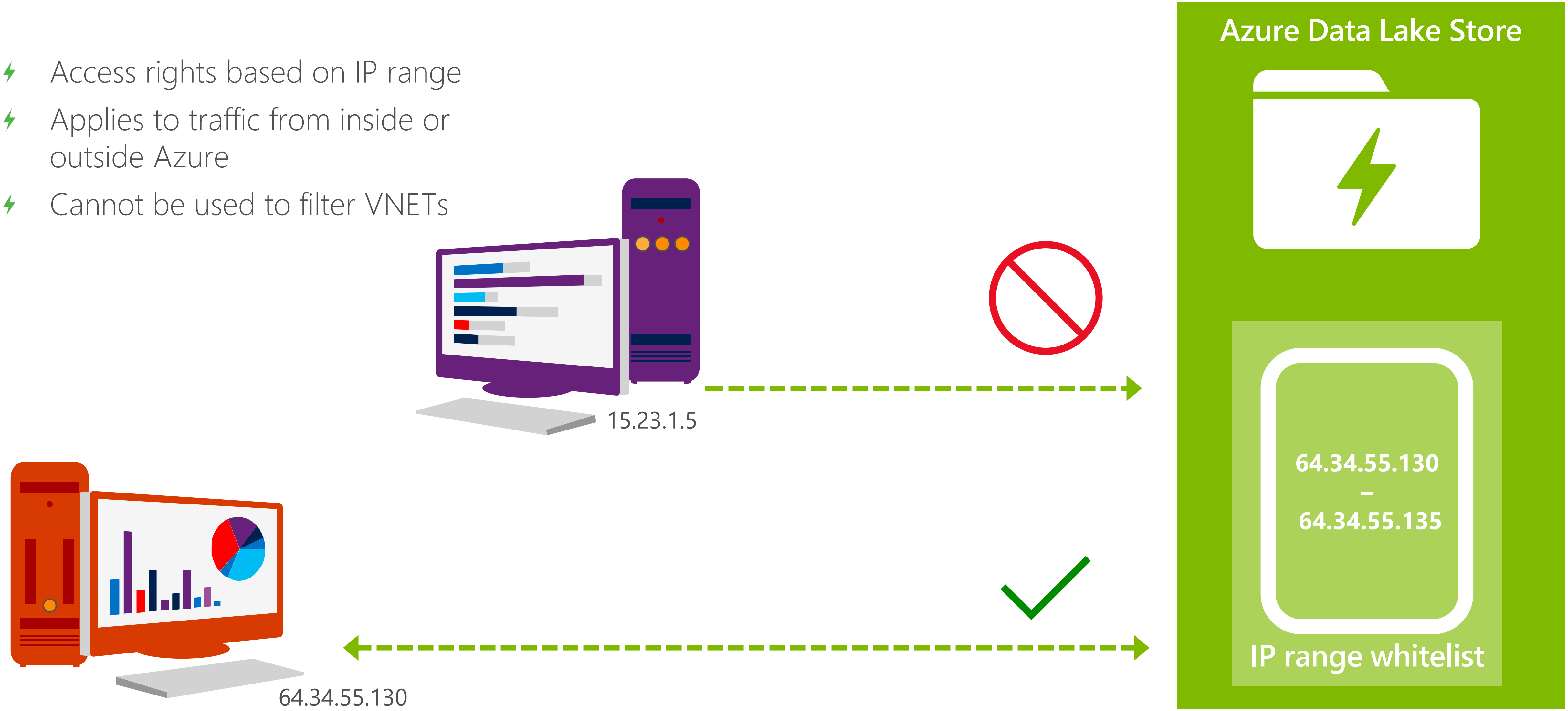
POSIX-Style ACLs with full compatibility with HDFS/WebHDFS

- 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



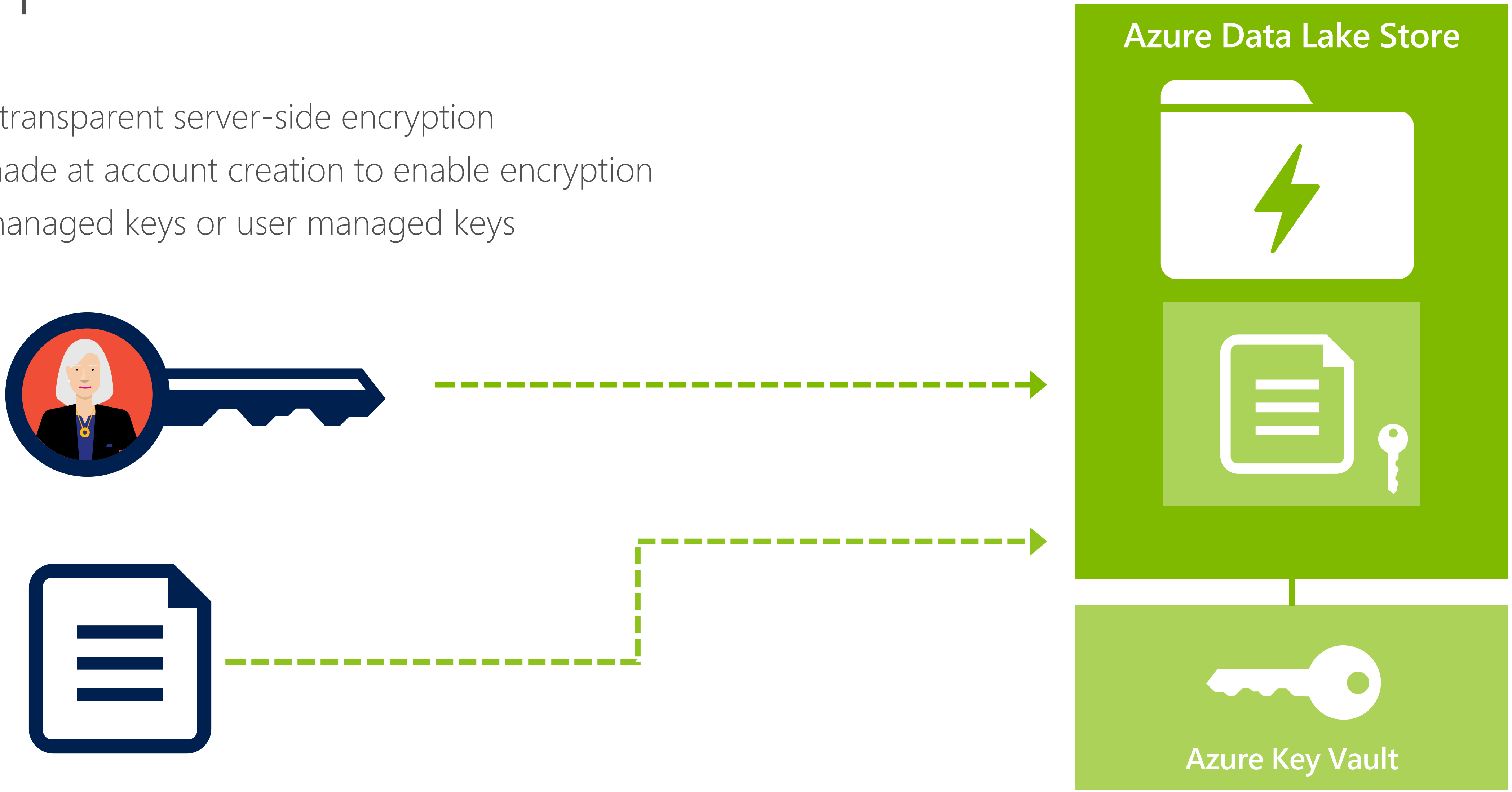
IP address ACLs

- 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*

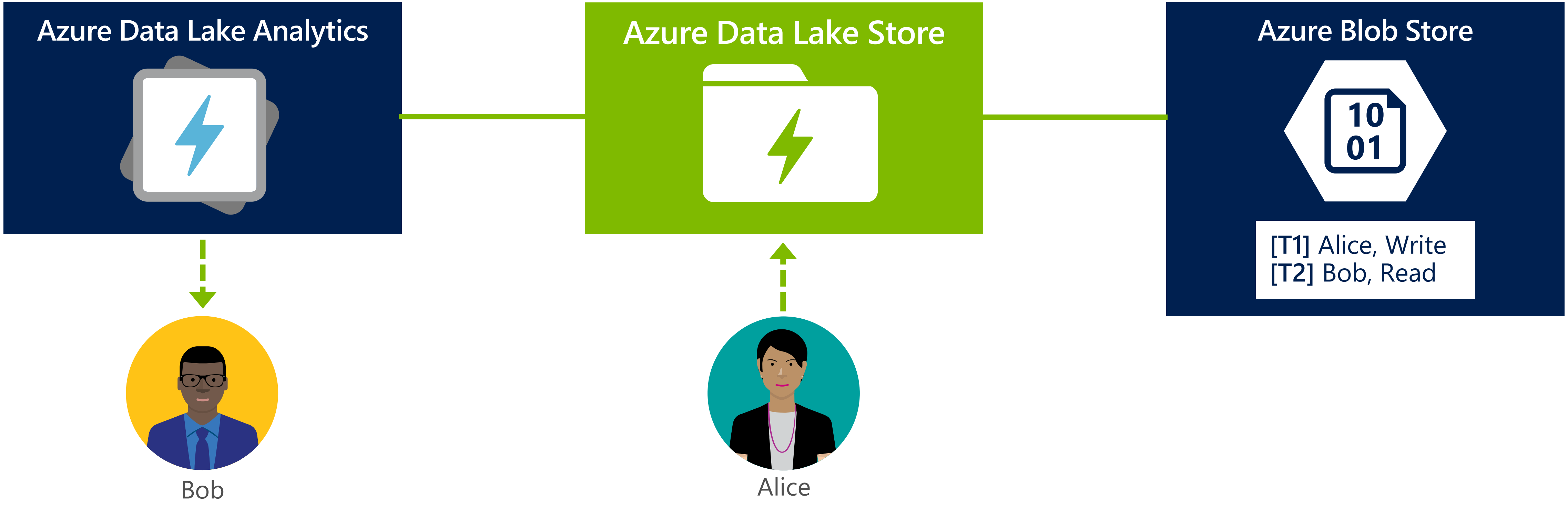
- Provides transparent server-side encryption
- Choice made at account creation to enable encryption
- Service managed keys or user managed keys



* In Private Preview

Audit logs for data access

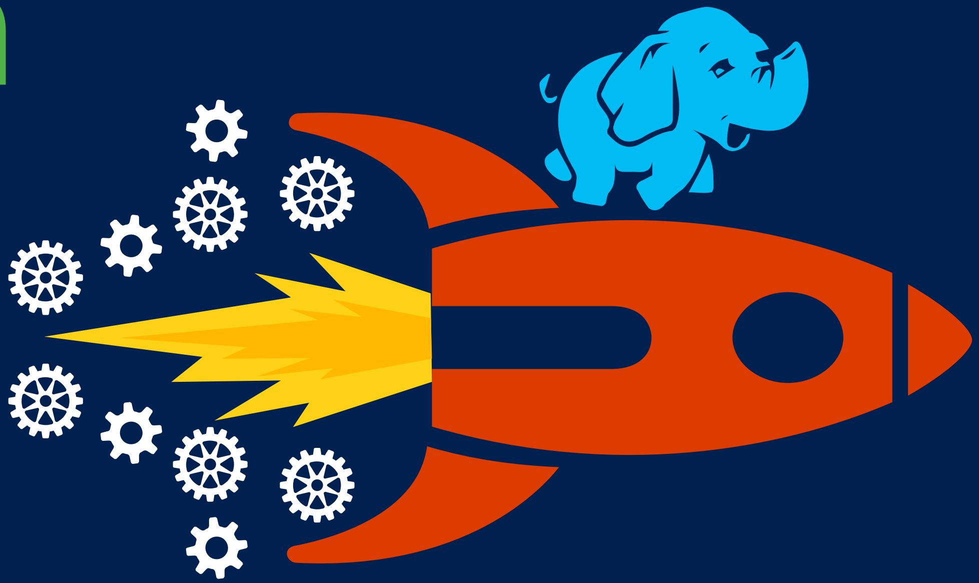
- Logs are available in JSON format
- Sample U-SQL scripts are available on GitHub to-read logs
- Enhancement to logs will continue through GA



[T1] Alice, Write
[T2] Bob, Read

ADL Store

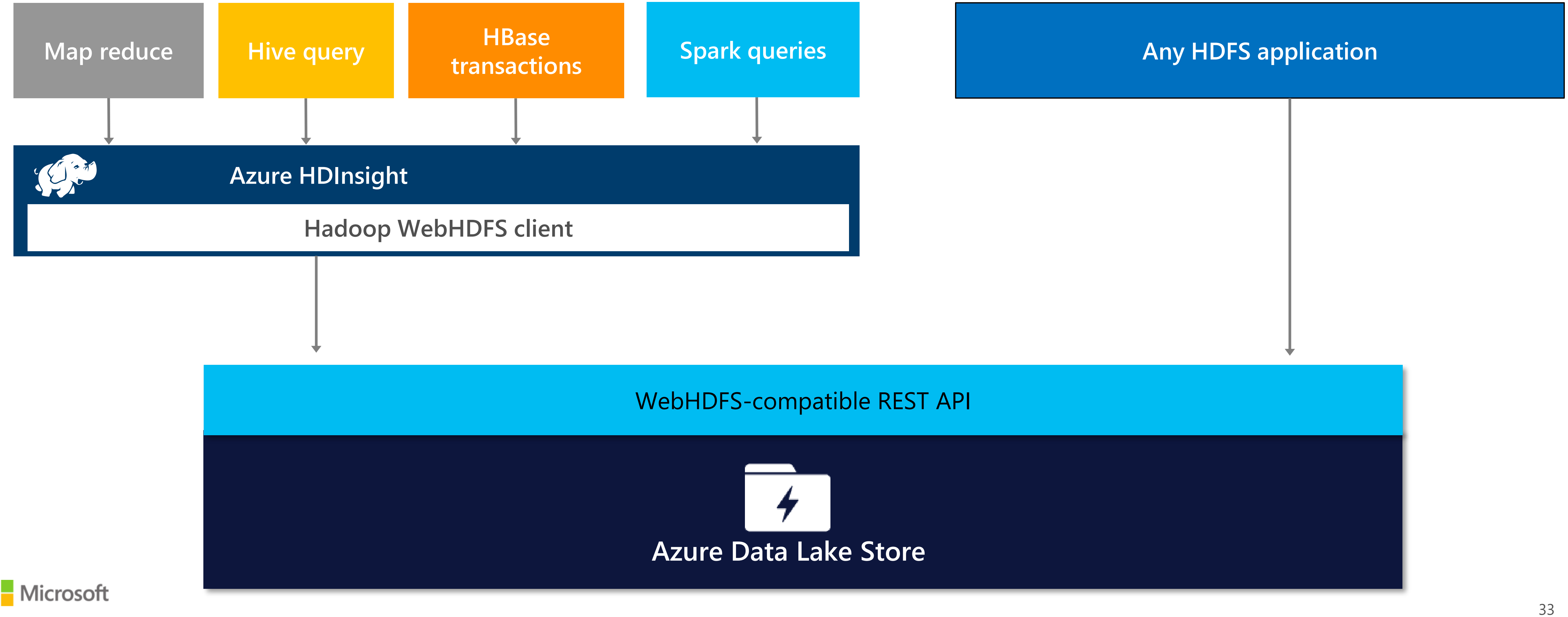
Hadoop integration



Microsoft

ADL Store is HDFS-compatible

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



Map reduce

Hive query

HBase transactions

Spark queries

Any HDFS application

Azure HDInsight

Hadoop WebHDFS client

WebHDFS-compatible REST API

Azure Data Lake Store

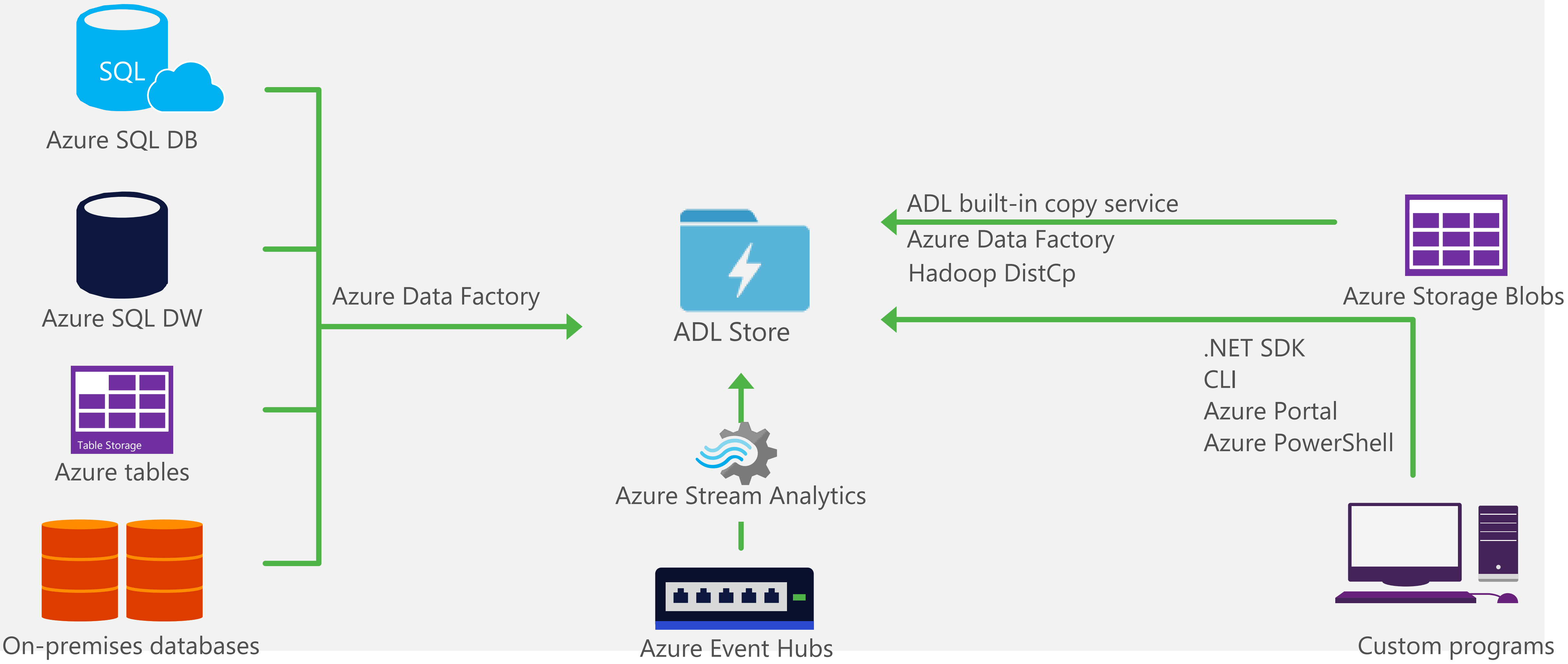
Microsoft

ADL Store: ingress and egress

Microsoft

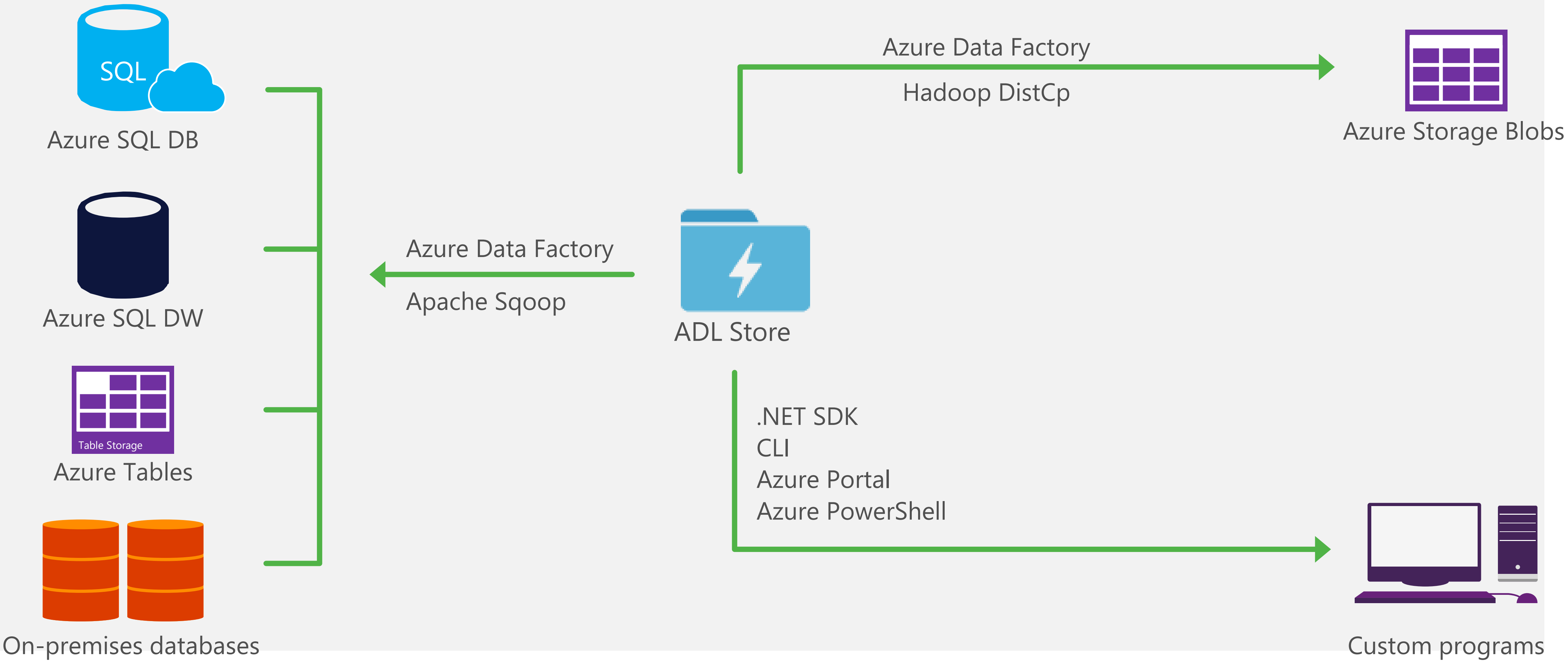
ADL Store: Ingress

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



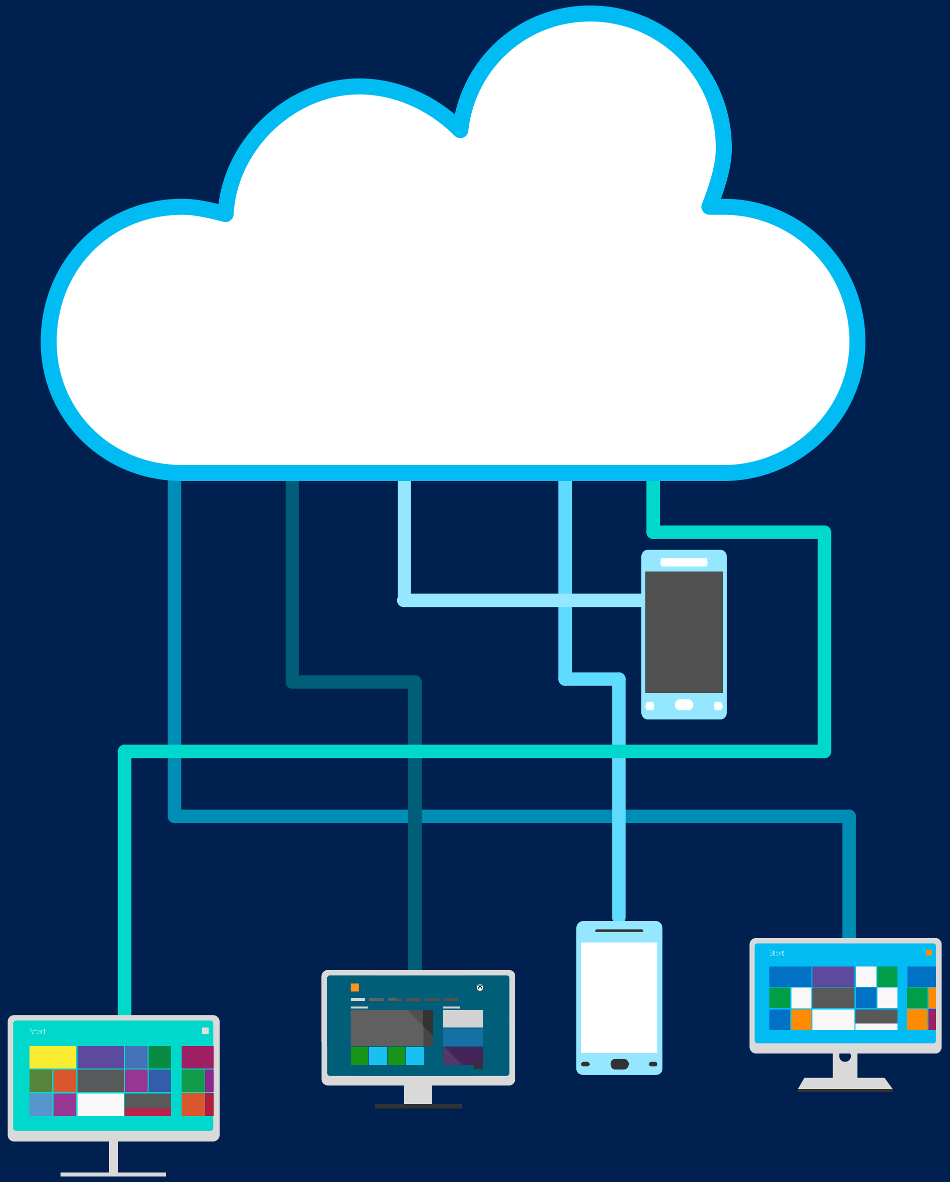
ADL Store: Egress

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



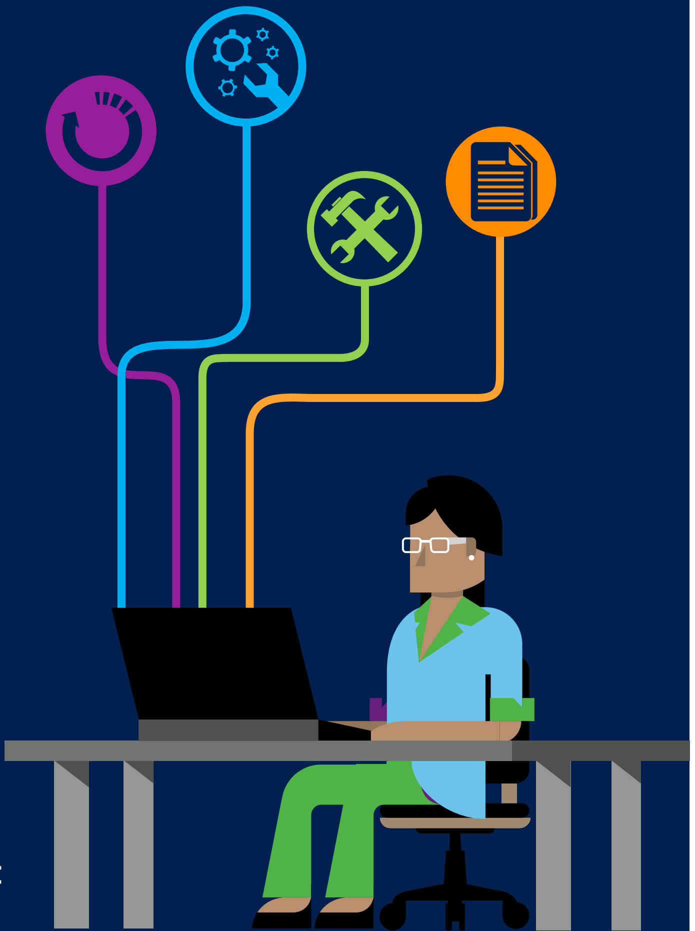
Microsoft

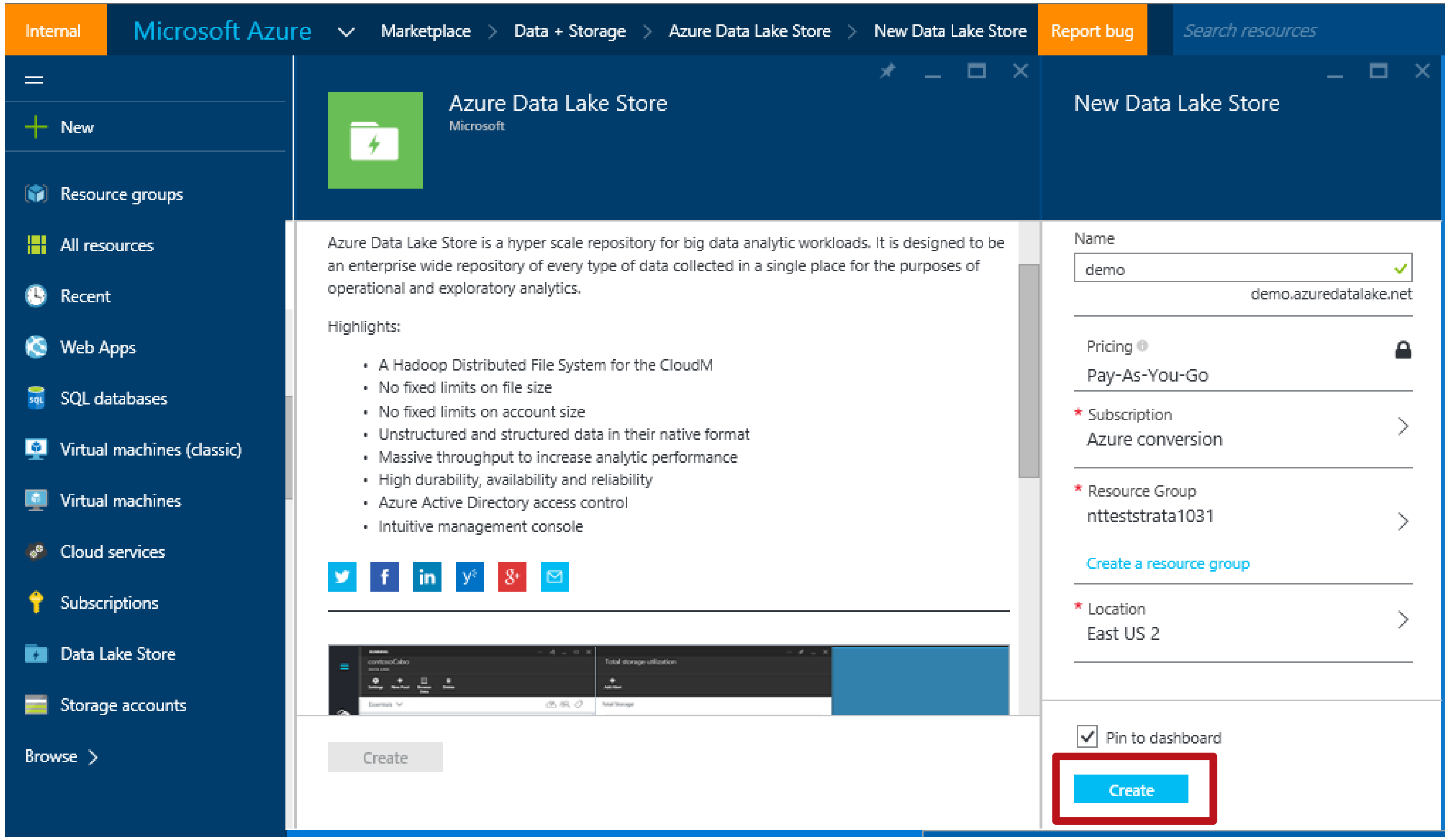
ADL Store: Azure Portal integration



Microsoft

Creating a new ADL Store





Azure Data Lake Store
Microsoft

Azure Data Lake Store is a hyper scale repository for big data analytic workloads. It is designed to be an enterprise wide repository of every type of data collected in a single place for the purposes of operational and exploratory analytics.

Highlights:

- A Hadoop Distributed File System for the CloudM
- No fixed limits on file size
- No fixed limits on account size
- Unstructured and structured data in their native format
- Massive throughput to increase analytic performance
- High durability, availability and reliability
- Azure Active Directory access control
- Intuitive management console

New Data Lake Store

Name
demo ✓
demo.azuredatalake.net

Pricing
Pay-As-You-Go

* Subscription
Azure conversion

* Resource Group
ntteststrata1031

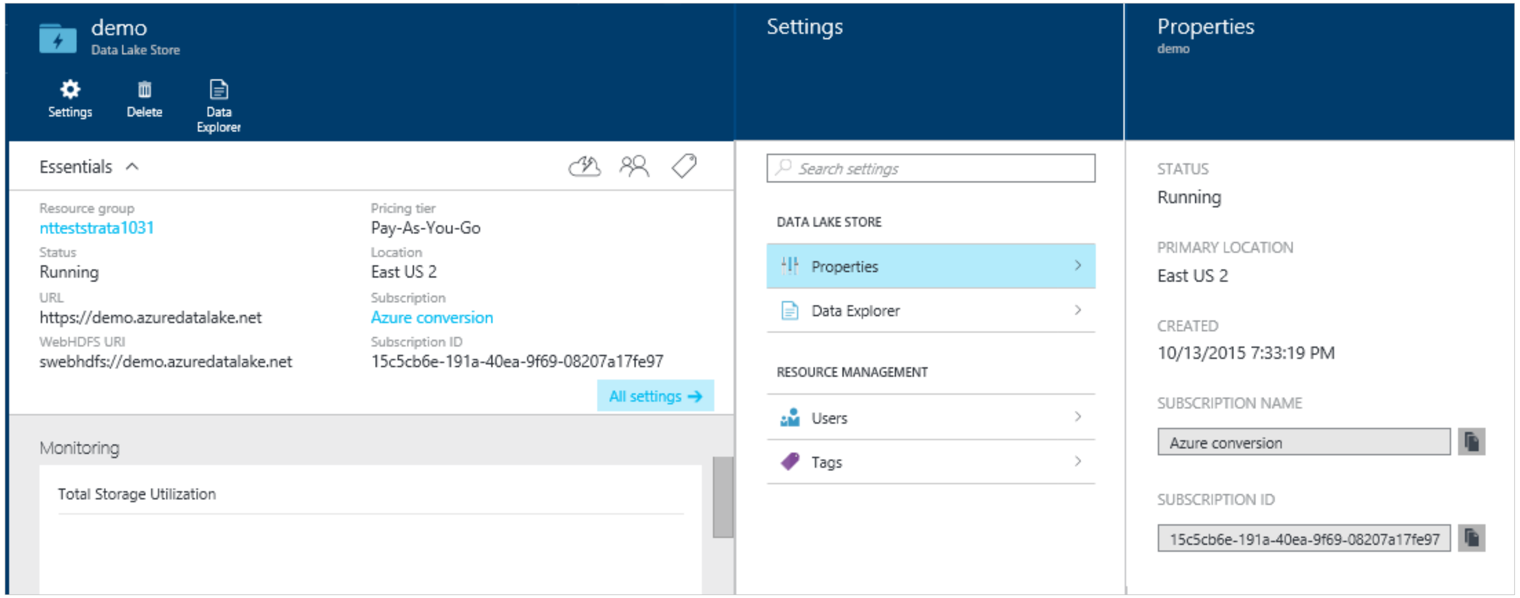
Create a resource group

* Location
East US 2

☑ Pin to dashboard

Create

ADL Store: Properties



Resource group
nttestStrata1031

Status
Running

URL
https://demo.azuredatalake.net

WebHDFS URI
swebhdfs://demo.azuredatalake.net

Pricing tier
Pay-As-You-Go

Location
East US 2

Subscription
Azure conversion

Subscription ID
15c5cb6e-191a-40ea-9f69-08207a17fe97

STATUS
Running

PRIMARY LOCATION
East US 2

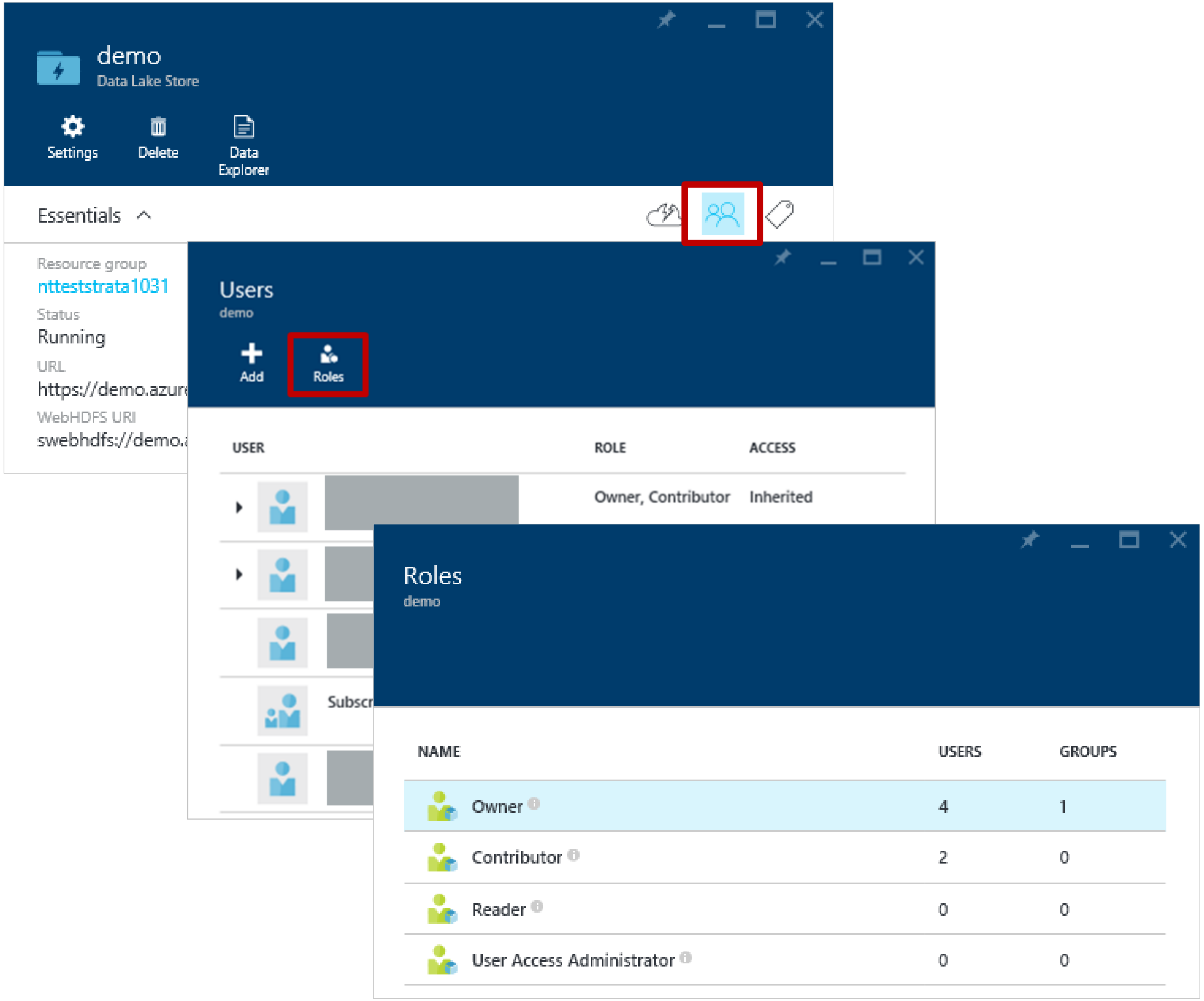
CREATED
10/13/2015 7:33:19 PM

SUBSCRIPTION NAME
Azure conversion

SUBSCRIPTION ID
15c5cb6e-191a-40ea-9f69-08207a17fe97

Microsoft

Viewing Users and their Roles & Privileges



Resource group
nttestrata1031

Status
Running

URL
https://demo.azur

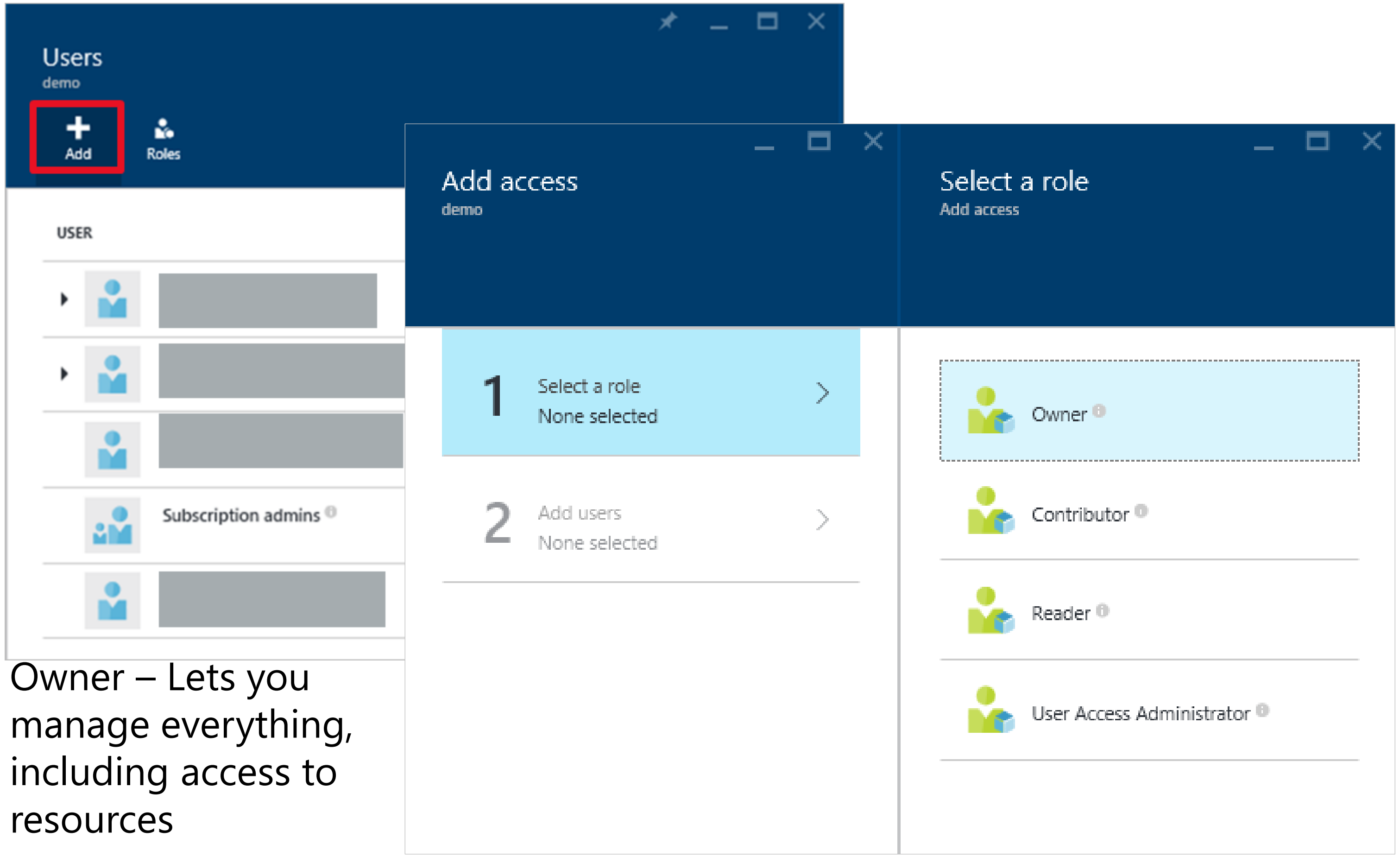
WebHDFS URI
swebhdfs://demo.a

USER	ROLE	ACCESS
[redacted]	Owner, Contributor	Inherited
[redacted]		
[redacted]		
Subscr		
[redacted]		

NAME	USERS	GROUPS
Owner	4	1
Contributor	2	0
Reader	0	0
User Access Administrator	0	0

Adding Users

A new user can be added in the role of Owner, Contributor, Reader or User Access Administrator



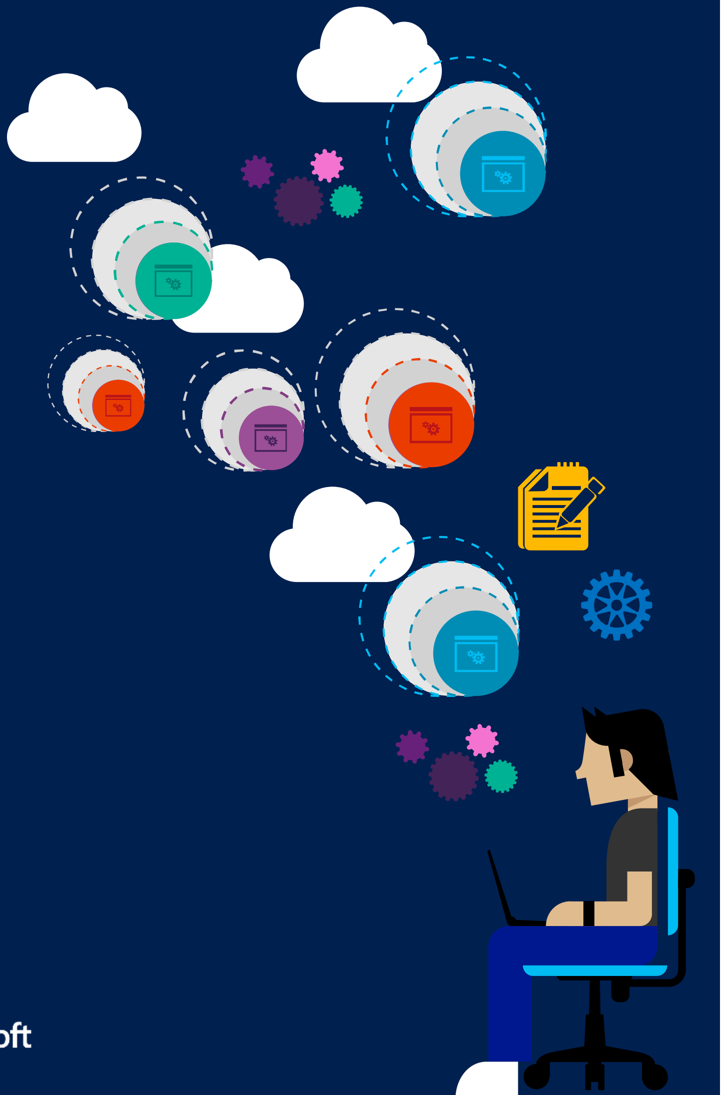
Owner – Lets you manage everything, including access to resources

Contributor – Lets you manage everything, except access to resources

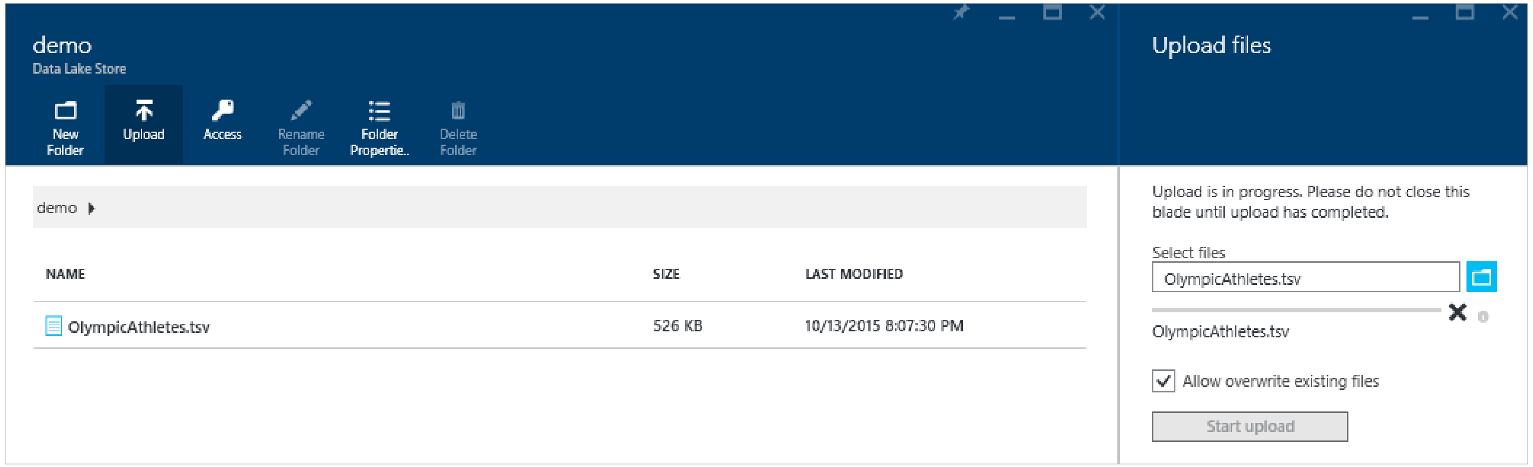
Reader – Lets you view everything, but not make changes

User Access Administrator – Lets you manage user access to Azure resources

File Upload



Azure Portal lets you upload files directly to ADL Store



NAME	SIZE	LAST MODIFIED
OlympicAthletes.tsv	526 KB	10/13/2015 8:07:30 PM

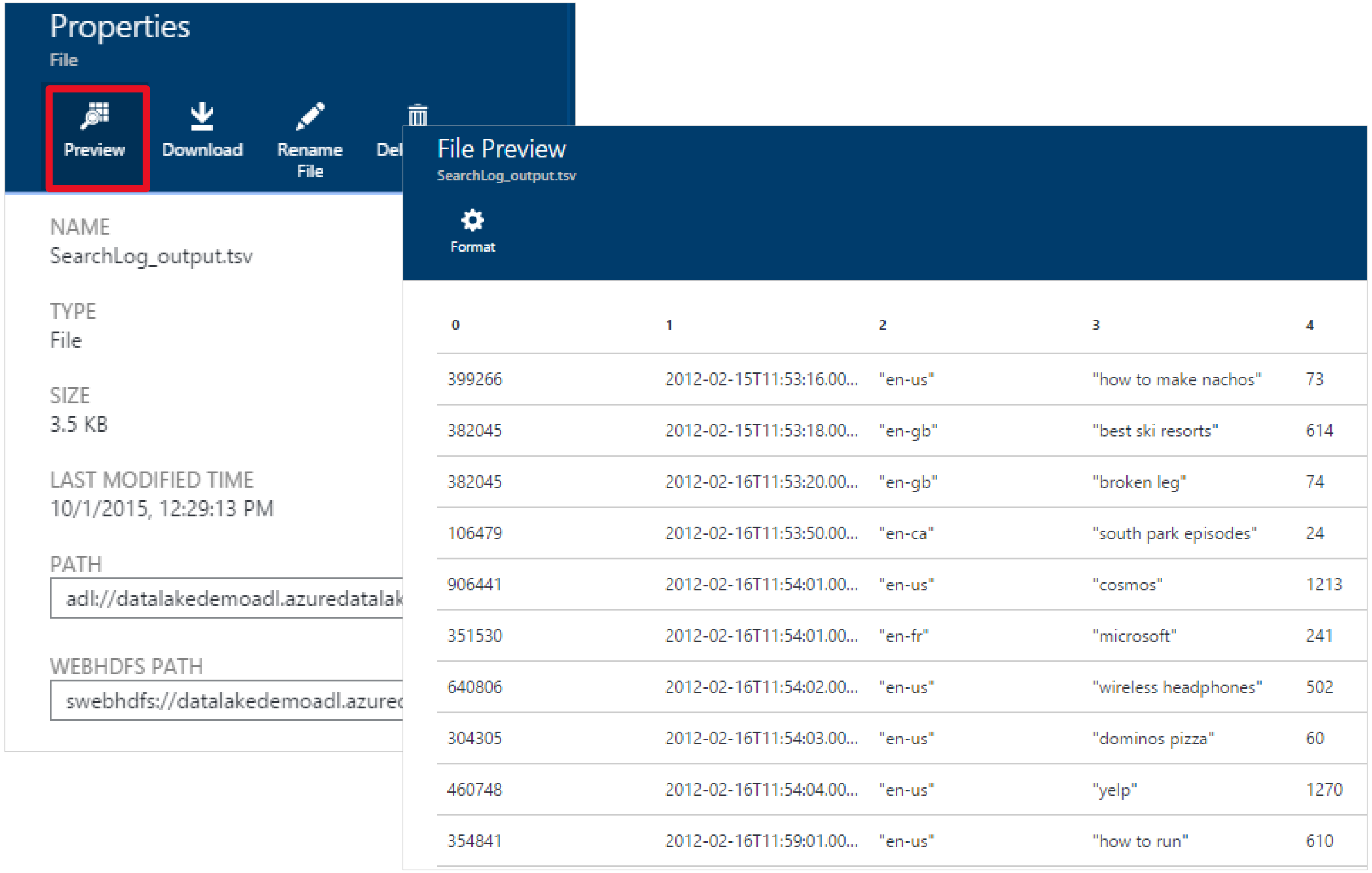
Upload is in progress. Please do not close this blade until upload has completed.

Select files: OlympicAthletes.tsv

OlympicAthletes.tsv

☑ Allow overwrite existing files

File Preview



NAME
SearchLog_output.tsv

TYPE
File

SIZE
3.5 KB

LAST MODIFIED TIME
10/1/2015, 12:29:13 PM

PATH
adl://datalakedemoadl.azuredatalak

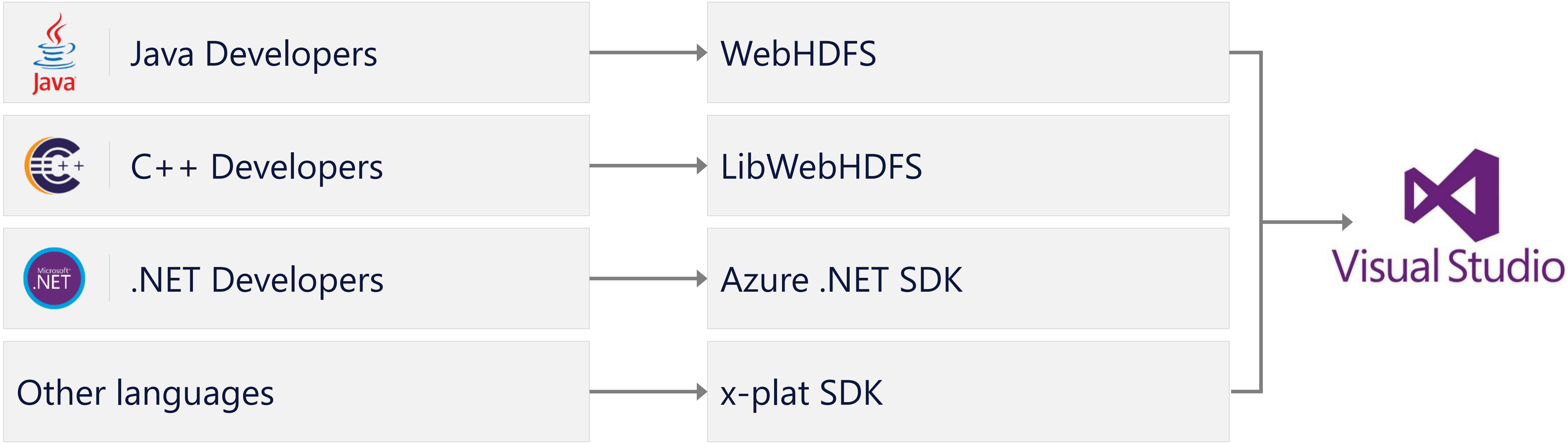
WEBHDFS PATH
swebhdfs://datalakedemoadl.azurec

0	1	2	3	4
399266	2012-02-15T11:53:16.00...	"en-us"	"how to make nachos"	73
382045	2012-02-15T11:53:18.00...	"en-gb"	"best ski resorts"	614
382045	2012-02-16T11:53:20.00...	"en-gb"	"broken leg"	74
106479	2012-02-16T11:53:50.00...	"en-ca"	"south park episodes"	24
906441	2012-02-16T11:54:01.00...	"en-us"	"cosmos"	1213
351530	2012-02-16T11:54:01.00...	"en-fr"	"microsoft"	241
640806	2012-02-16T11:54:02.00...	"en-us"	"wireless headphones"	502
304305	2012-02-16T11:54:03.00...	"en-us"	"dominos pizza"	60
460748	2012-02-16T11:54:04.00...	"en-us"	"yelp"	1270
354841	2012-02-16T11:59:01.00...	"en-us"	"how to run"	610

- Input and output files can be previewed directly in the portal without having to download them.
- The preview shows the first few rows.
- Column numbers are automatically assigned
- Understands CSV and TSV formats.

App Development – Languages and Tools

Azure Data Lake Store supports multiple languages for application development



Java Developers → WebHDFS

C++ Developers → LibWebHDFS

.NET Developers → Azure .NET SDK

Other languages → x-plat SDK

→ Visual Studio

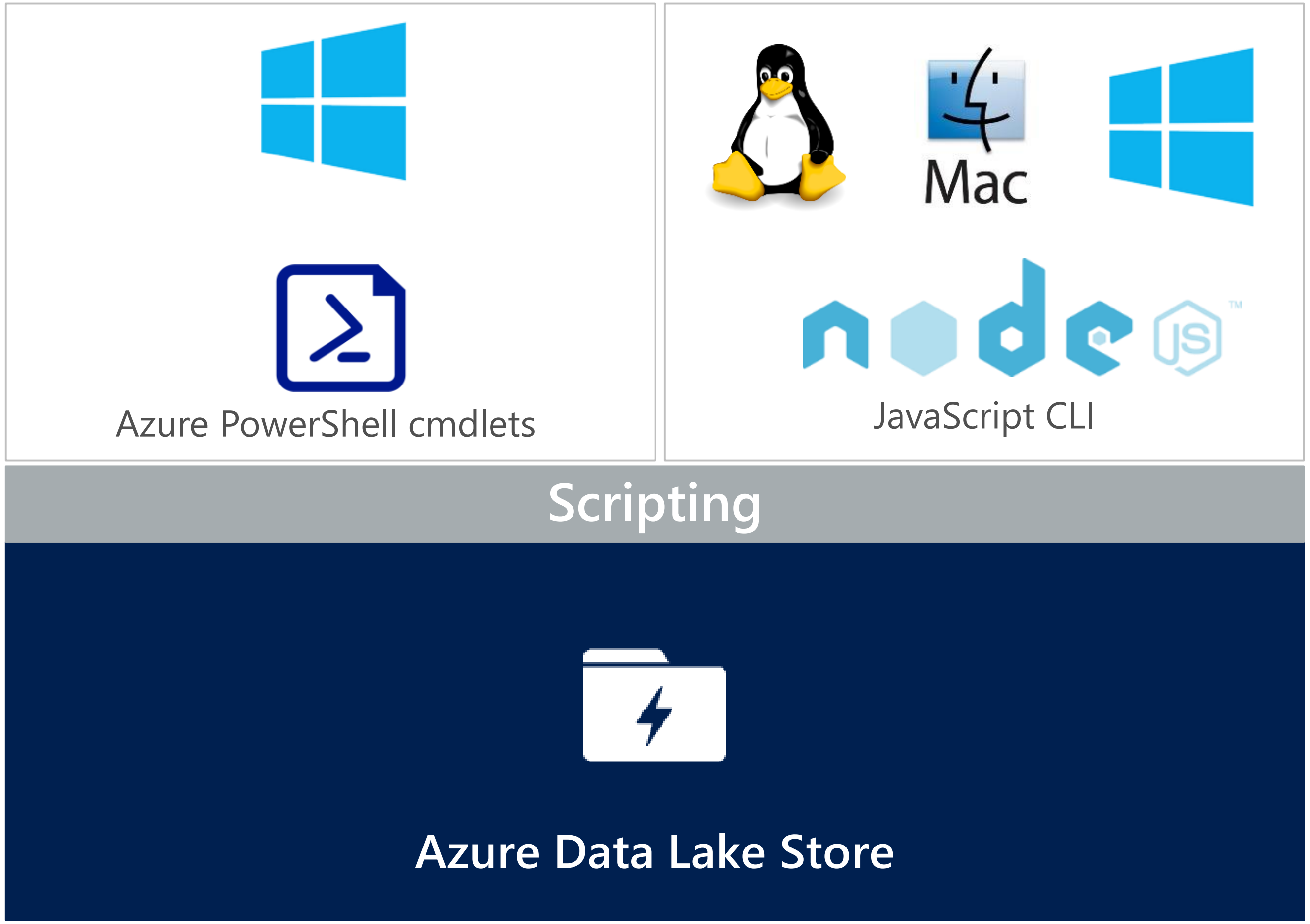
Note: If you are using Hadoop (Map Reduce programs or Hive or HBase) or Spark, then you will not be programming directly to the Azure Data Lake Store as they all will transparently access Azure Data Lake Store under the covers.

Developing scripting applications

Provides native Windows and cross-platform (Mac, Linux) scripting experience

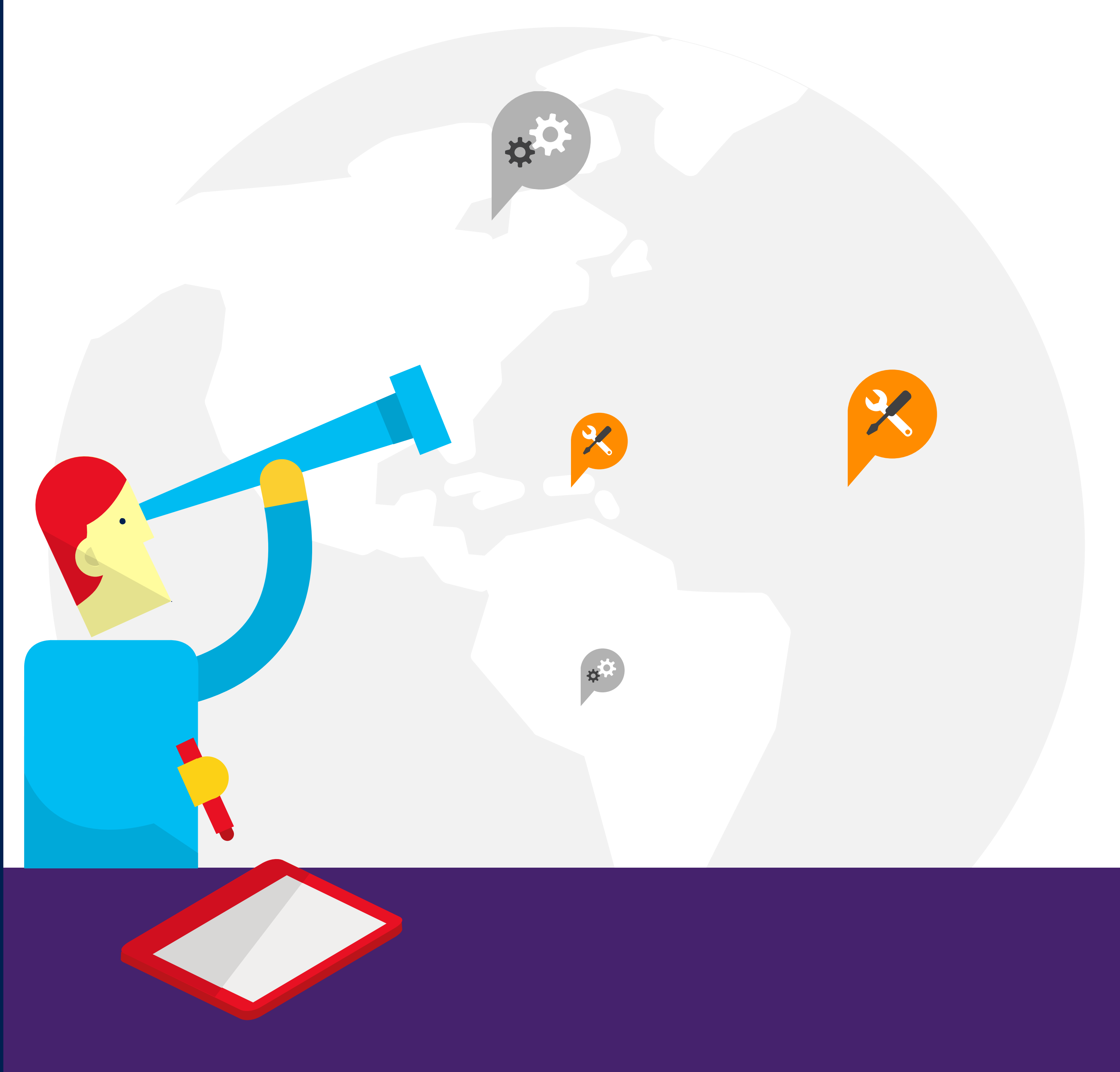
Scripting operations include

- Create new directories
- Listing the contents of a directory
- Upload files to directory
- Delete files/directories
- Rename files/directories
- …



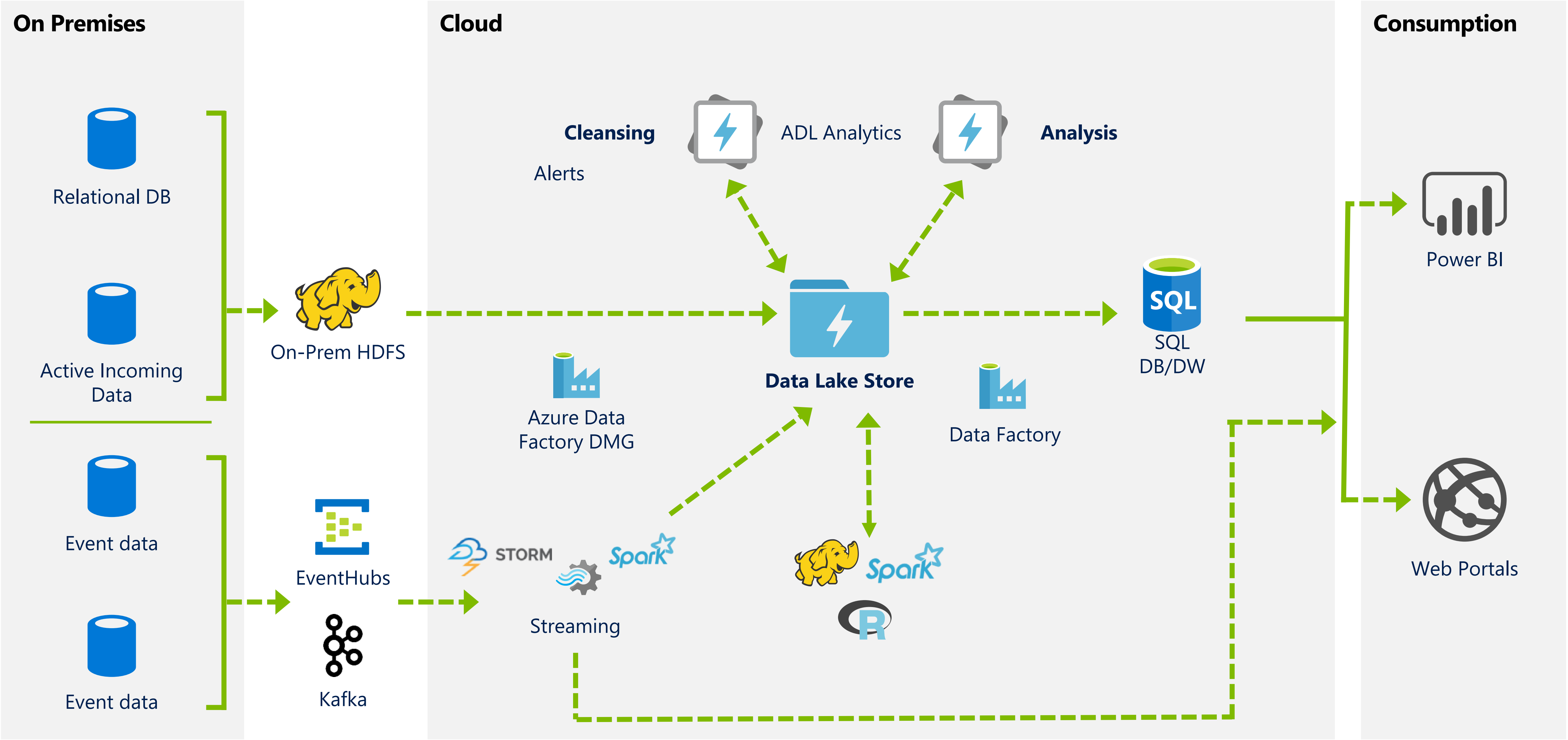
Implementation

Common customer patterns



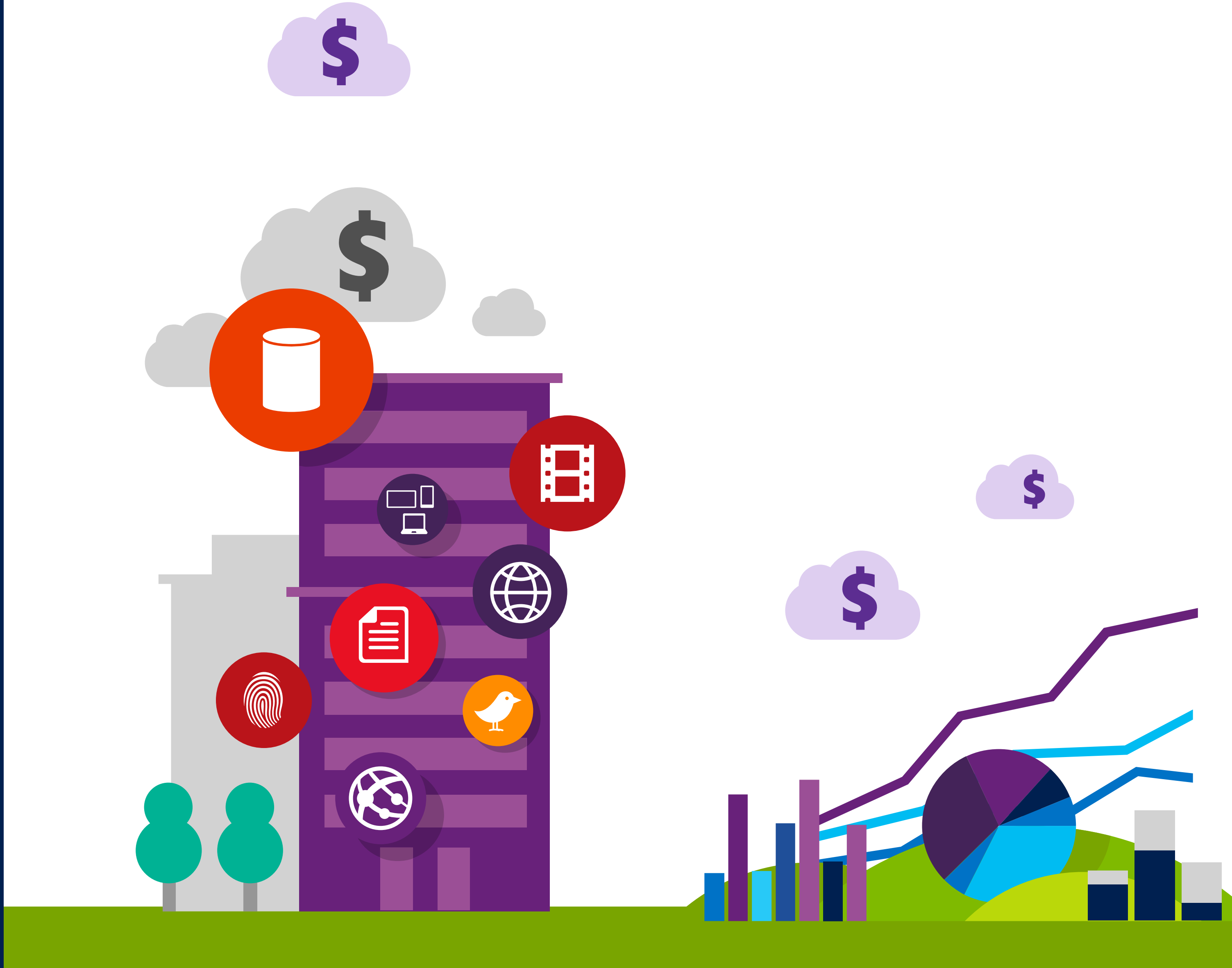
Microsoft

Lambda architecture



ADL 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

Get started today!

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

Microsoft