Azure Data Lake Overview

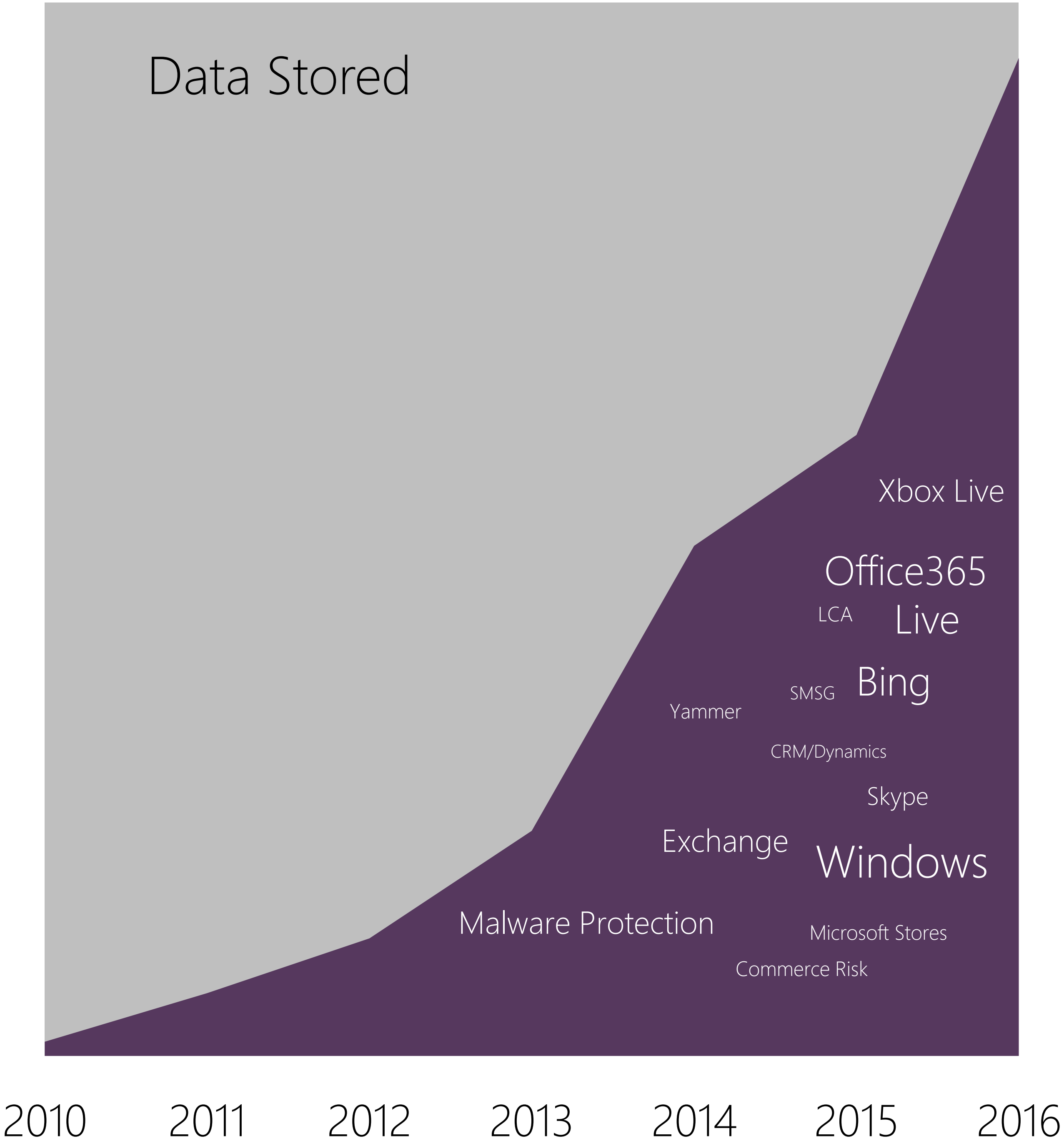
Mithun Prasad, PhD
Senior Program Manager @ Microsoft

Microsoft's internal Data Lake (Cosmos)

- A data lake for everyone to put their data
- Tools approachable by any developer
- Batch, Interactive, Streaming, ML
- Used across Office, Xbox Live, Azure, Windows, Bing, Skype, etc…

By the numbers

- Exabytes of data under management
- 100Ks of Physical Servers
- 100Ks of Batch Jobs
- Millions of Interactive Queries
- Huge Streaming Pipelines
- 10K+ Developers running diverse workloads and scenarios



Reflections on Data

Collect the data first

We couldn’t have predicted the how we would get value from the data when we started collecting.

The Power of Sharing

A side effect of having a unified platform, with a consistent security model. Tools allow me to shape and join all data, while maintaining security, auditing and compliance for key data sets.

Data Virality

Value large datasets helps bootstrap the entire company into using big data.

Visibility & Control more important than ever

Ever growing needs for Auditing, Compliance, Data provenance, Regulatory

Reflections on Engineering

Data Agility

Build systems and data pipelines that allow producers and consumers of data to innovate rapidly

Changing Skillsets

Ramp up hiring of people with experience with data science, Machine Learning, etc.

.

About Creating Value; Not Minimizing Cost

We are spending more to deliver more value for the business. Have become much more mature about spending and efficiency.

Azure Data Lake

Harness all your data.

Bring intelligent actions to your business.

Microsoft

1

Any Data

2

Enterprise

3

Developers

The 3 Azure Data Lake Services

Analytics

Big data queries as a service

Store

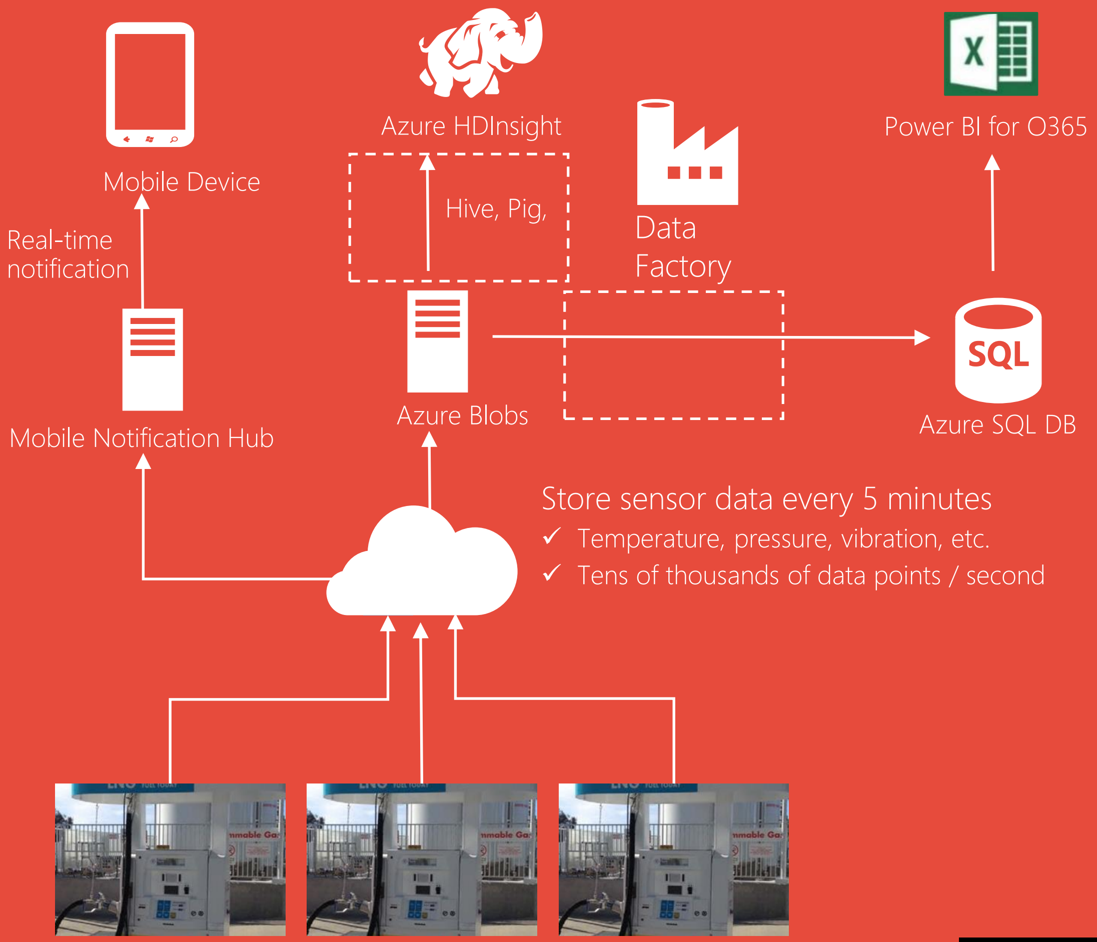
Hyper-scale Storage optimized for analytics

HDInsight

Clusters as a service

Scenarios

Rockwell Automation has partnered with one of the six oil and gas super majors to build unmanned internet-connected gas dispensers. Each dispenser emits real-time management metrics allowing them to detect anomalies and predict when proactive maintenance needs to occur.



Store sensor data every 5 minutes

- ✓ Temperature, pressure, vibration, etc.
- ✓ Tens of thousands of data points / second

Microsoft

Rockwell Automation

One of the leaders in the development and management of renewable energy infrastructure and services needed to understand data coming from their wind turbines/wind farms in an Internet of Things (IoT) scenario.

- 100s of windfarms across the globe
- Each windfarm has 100+ turbines
- Each turbine generates 10 data points every 25 milliseconds.

Initial goal

Provide consumption related analytics to their customers (power companies)

What else could they do with all that data?

Predictive maintenance

How?

Event Hub, Azure Storage, HDInsight

Azure SQL DB, Excel reporting

Microsoft

Microsoft Device Telemetry Pipeline

Windows 10, XBOX, Services, …

1 Billion

Total Devices

Hundreds of Billions

Events Processed Daily

Hundreds of Terabytes

Raw Data Ingested Daily

Real-time and Near-real-time Analysis of…

- Game Achievements
- Device Crash Logs
- Bluescreen crashdump
- Etc.

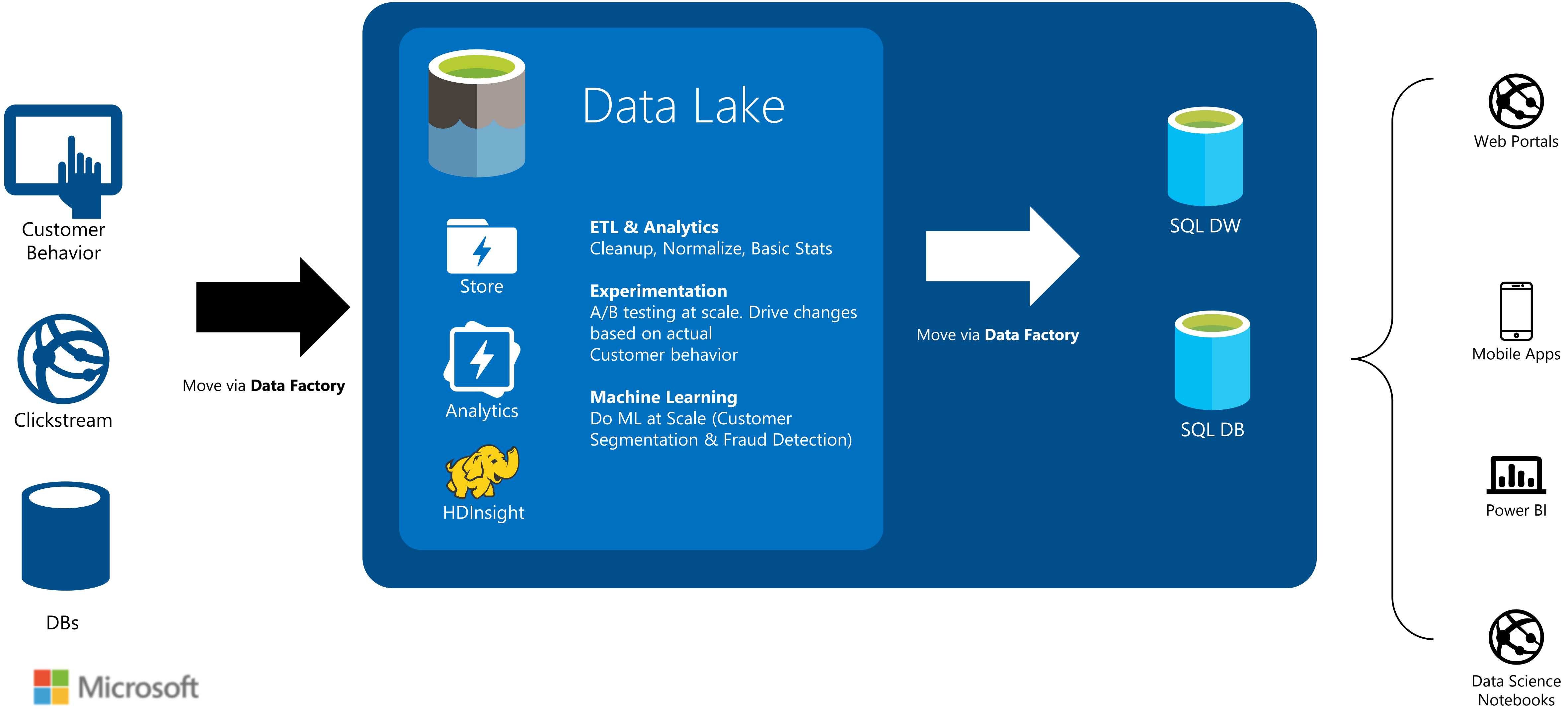
Critical for “Windows Service”

Data Lake Analytics Scenario

ON PREMISES

CLOUD

CONSUMPTION



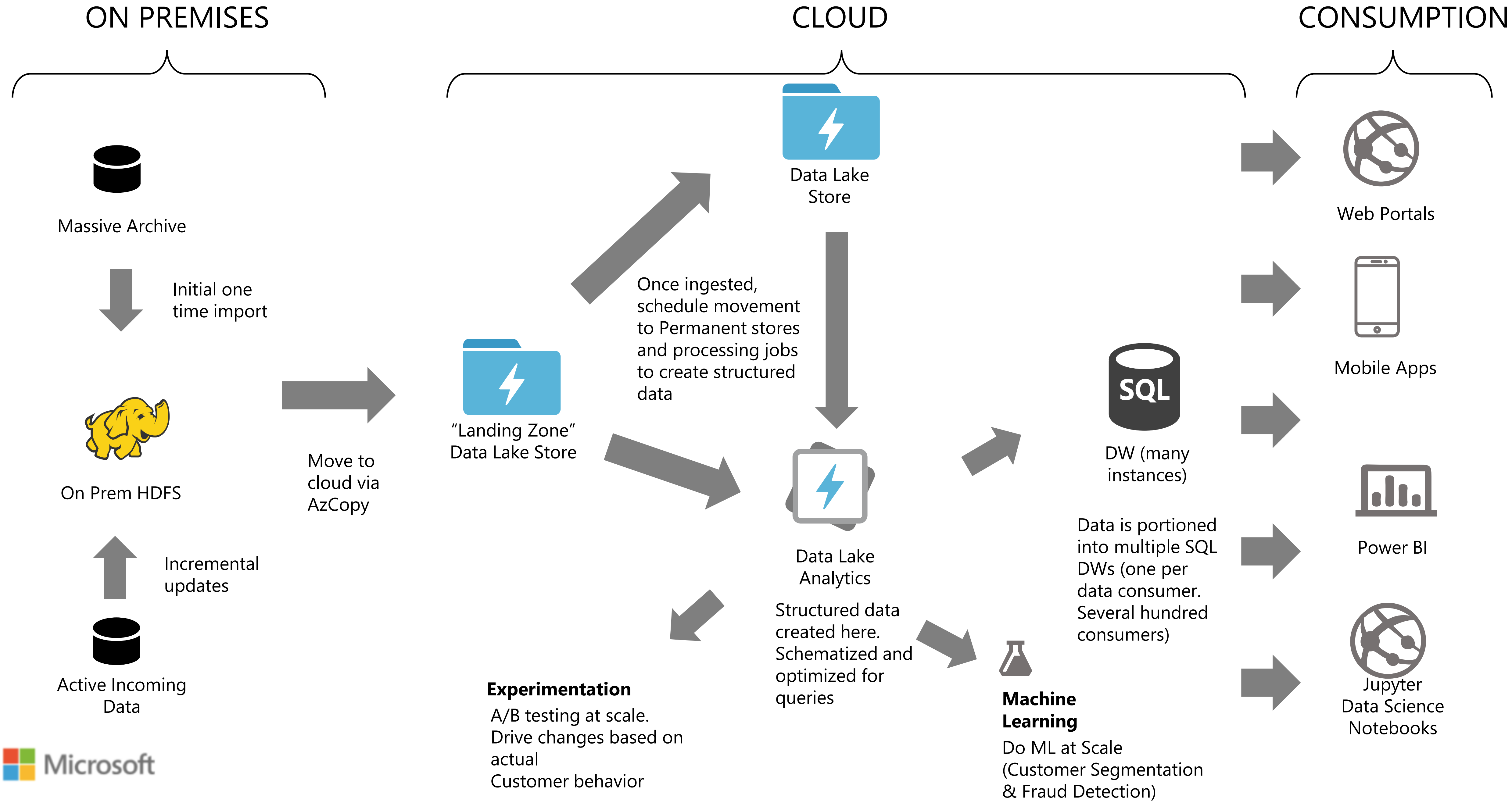
ETL & Analytics
Cleanup, Normalize, Basic Stats

Experimentation
A/B testing at scale. Drive changes based on actual Customer behavior

Machine Learning
Do ML at Scale (Customer Segmentation & Fraud Detection)

Microsoft

Real World Scenario with Azure Data Lake (Retail)



ON PREMISES

Massive Archive

Initial one time import

On Prem HDFS

Incremental updates

Active Incoming Data

Move to cloud via AzCopy

CLOUD

“Landing Zone” Data Lake Store

Once ingested, schedule movement to Permanent stores and processing jobs to create structured data

Data Lake Store

Data Lake Analytics

Structured data created here. Schematized and optimized for queries

Experimentation

A/B testing at scale. Drive changes based on actual Customer behavior

SQL

DW (many instances)

Data is portioned into multiple SQL DWs (one per data consumer. Several hundred consumers)

Machine Learning

Do ML at Scale (Customer Segmentation & Fraud Detection)

CONSUMPTION

Web Portals

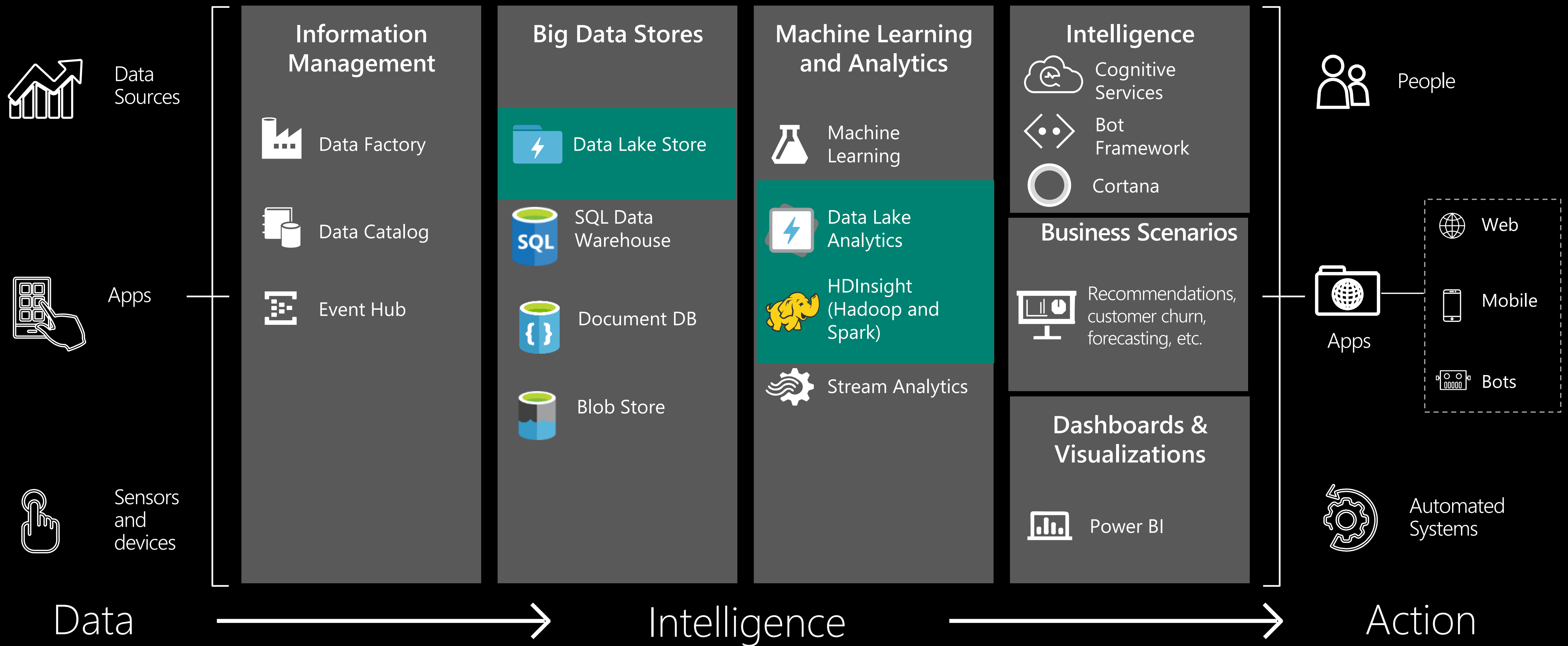
Mobile Apps

Power BI

Jupyter Data Science Notebooks

Cortana Intelligence Suite

Transform data into intelligent actions and predictions



Data Sources

Apps

Sensors and devices

Information Management

Data Factory

Data Catalog

Event Hub

Big Data Stores

Data Lake Store

SQL Data Warehouse

Document DB

Blob Store

Machine Learning and Analytics

Machine Learning

Data Lake Analytics

HDInsight (Hadoop and Spark)

Stream Analytics

Intelligence

Cognitive Services

Bot Framework

Cortana

Business Scenarios

Recommendations, customer churn, forecasting, etc.

Dashboards & Visualizations

Power BI

People

Apps

Web

Mobile

Bots

Automated Systems

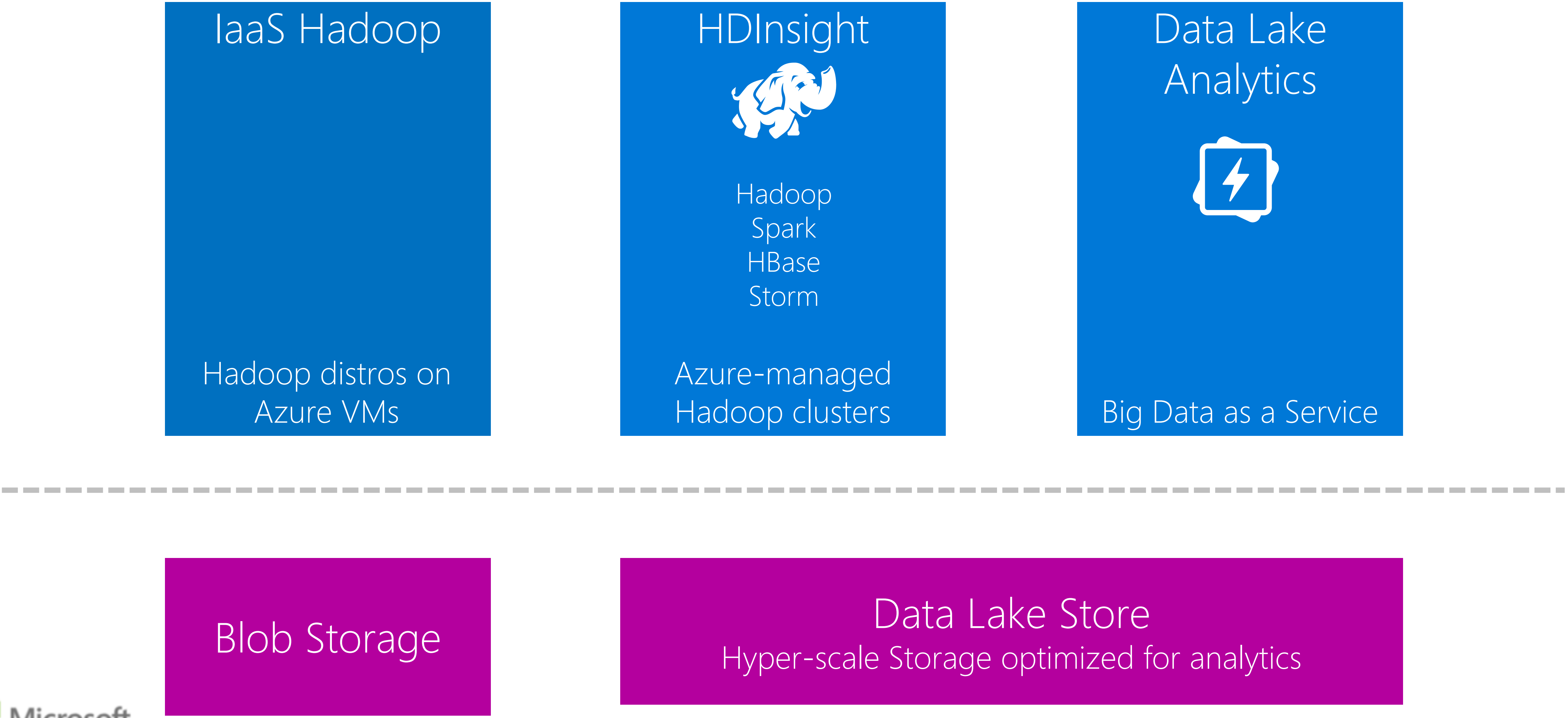
Data → Intelligence → Action

Microsoft

Hadoop & Big Data Azure

Microsoft

Big Data in Azure



IaaS Hadoop

Hadoop distros on Azure VMs

HDInsight

Hadoop
Spark
HBase
Storm

Azure-managed Hadoop clusters

Data Lake Analytics

Big Data as a Service

Blob Storage

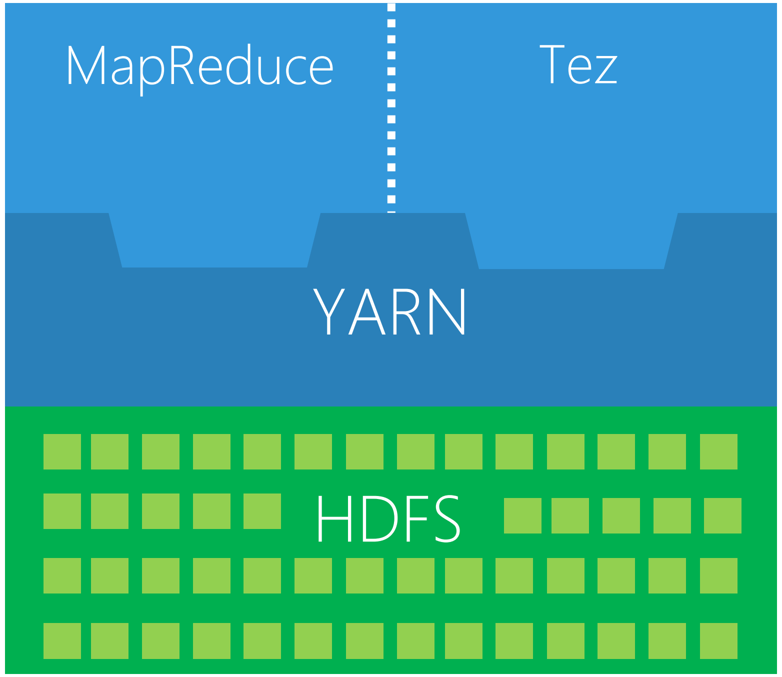
Data Lake Store
Hyper-scale Storage optimized for analytics

Apache Hadoop

A highly reliable, distributed and parallel programming framework for analyzing big data

- Open source
- Java-based
- Runs on variety of hardware platforms, including clusters of commodity hardware
- Tolerant to failures of nodes, software components, network
- Scales with the cluster
- Rich ecosystem that supports SQL/NoSQL, Streaming, Real-time and Interactive applications.

Hadoop Core



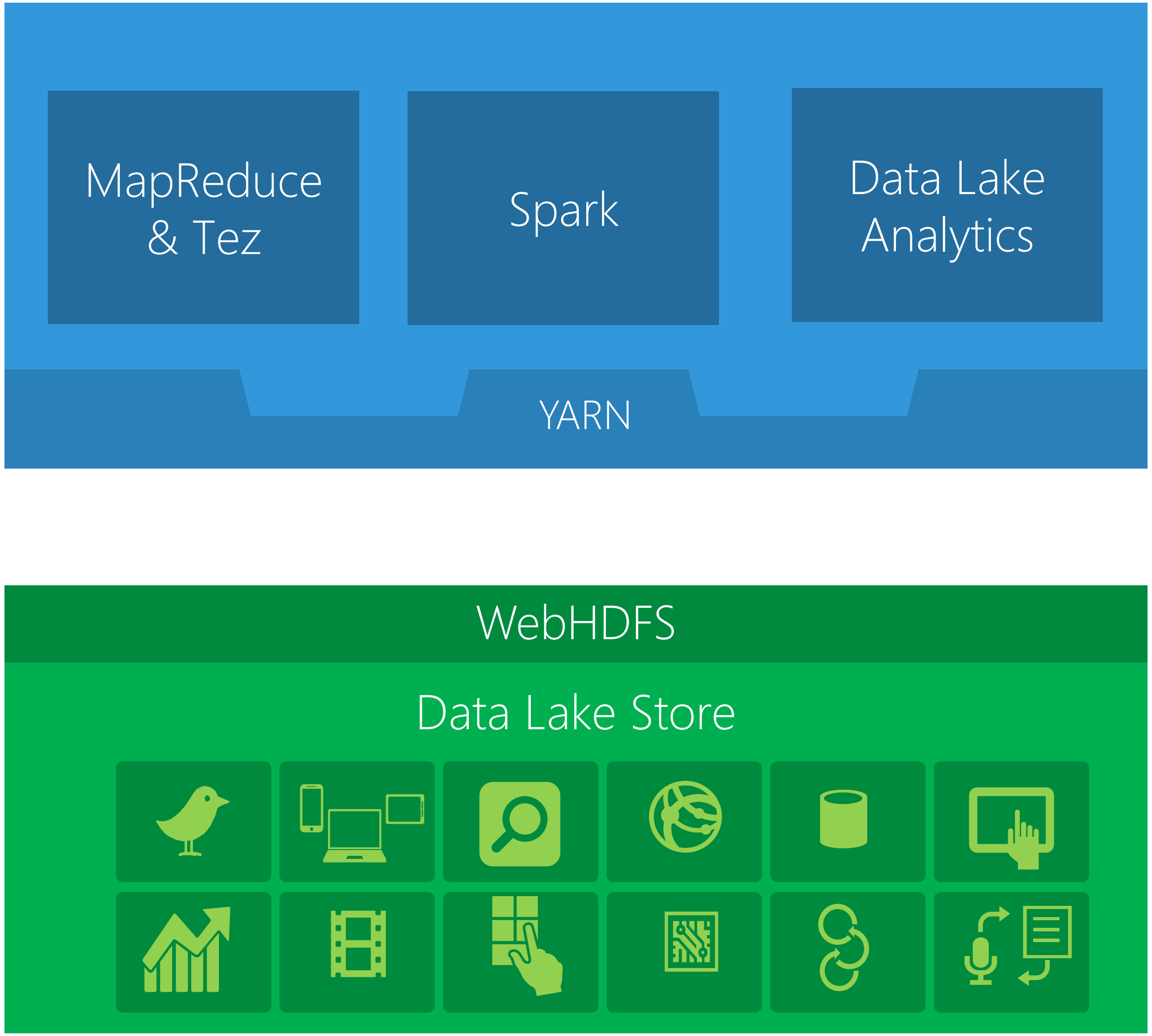
Develop programs based on **MapReduce**, **Tez**, etc. on top of **YARN**

YARN, a distributed resource manager that allocates and controls access to the resources of the cluster manager

HDFS - A scalable, reliable file system (HDFS)

Microsoft

Compute Workloads on YARN



MapReduce & Tez

Spark

Data Lake Analytics

YARN

WebHDFS

Data Lake Store

Azure Data Lake Store

Microsoft

Azure Data Lake Store

A No limits Data Lake that powers Big Data Analytics

The first cloud Data Lake for enterprises that is secure, massively scalable and built to the open HDFS standard. With no limits to the size of data and the ability to run massively parallel analytics, you can now unlock value from all your unstructured, semi-structured and structured data.

Petabyte size files and Trillions of objects

Scalable throughput for massively parallel analytics

HDFS for the Cloud

Always encrypted, Role-based Security & Auditing

Enterprise-grade Support

Petabyte size files and Trillions of objects

Reliable

- Automatically replicates your data
- Three copies within a single region
- Highly available

Unlimited Storage

- Unlimited account sizes
- Individual file sizes from gigabytes to petabytes
- No limits to scale

Scalable throughput for massively parallel analytics

Built for running large analytics systems that require massive throughput

Optimized for parallel I/O

Automatically optimizes for any throughput

HDFS for the Cloud

Built from the ground up as a Hadoop file system

HDI Cluster Types

- Hadoop
- Storm
- HBase (Future)
- Spark

Hadoop Distros

- Hortonworks (Future)
- Cloudera (Future)

Tools running in HDInsight Clusters

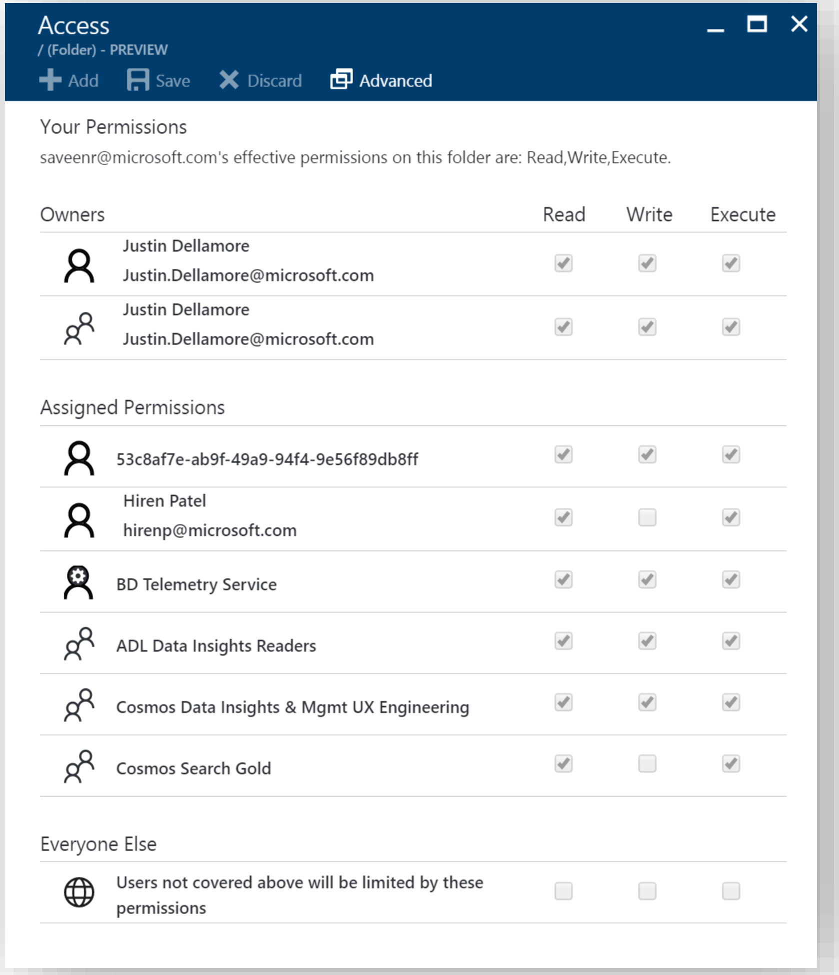
- Sqoop
- Distcp

Other

- Microsoft R Services
- Apache Hadoop (v2.8)

Always encrypted, Role-based Security & Auditing

- Role-based Access Control
- POSIX-compliant Access Control Lists (ACLs) on Files and Folders
- Integrated with Azure Active Directory
- Auditing for all operations. Audit logs that can be analysed with ADL U-SQL Scripts
- Transparent server-side encryption with Azure-managed (Azure Key Vault) and customer-managed keys



Your Permissions

saveenr@microsoft.com's effective permissions on this folder are: Read,Write,Execute.

Owners	Read	Write	Execute
Justin Dellamore Justin.Dellamore@microsoft.com	☑	☑	☑
Justin Dellamore Justin.Dellamore@microsoft.com	☑	☑	☑

Assigned Permissions	Read	Write	Execute
53c8af7e-ab9f-49a9-94f4-9e56f89db8ff	☑	☑	☑
Hiren Patel hirenp@microsoft.com	☑	□	☑
BD Telemetry Service	☑	☑	☑
ADL Data Insights Readers	☑	☑	☑
Cosmos Data Insights & Mgmt UX Engineering	☑	☑	☑
Cosmos Search Gold	☑	□	☑

Everyone Else	Read	Write	Execute
Users not covered above will be limited by these permissions	□	□	□

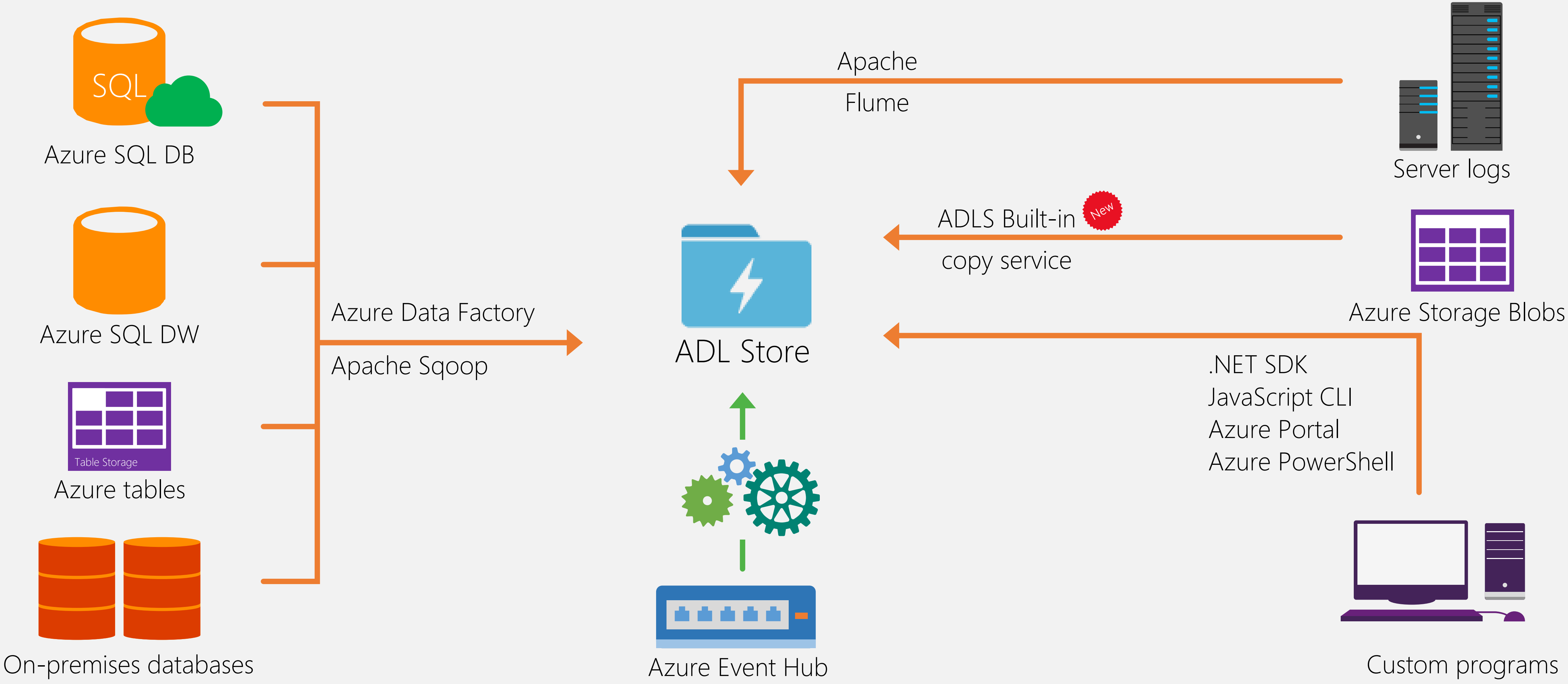
Data Lake Store vs Blob Storage

	Azure Data Lake Store	Azure Blob Storage
Purpose	Optimized for Analytics	General purpose bulk storage
Scenarios	Batch, Interactive, Streaming, ML	App backend, backup data, media storage for streaming
Units of Storage	Accounts / Folders / Files	Accounts / Containers / Blobs
Structure	Hierarchical File System	Flat namespace
Supports WebHDFS	Yes	No
Billing	Pay for data stored and for I/O	Pay for data stored and for I/O
Region Availability	US (Other regions coming)	All Azure Regions
Authentication	Azure Active Directory	Access keys
Authorization	POSIX ACLs on Files and Folders	Access Keys
Server-side Encryption	Yes	Yes

Microsoft

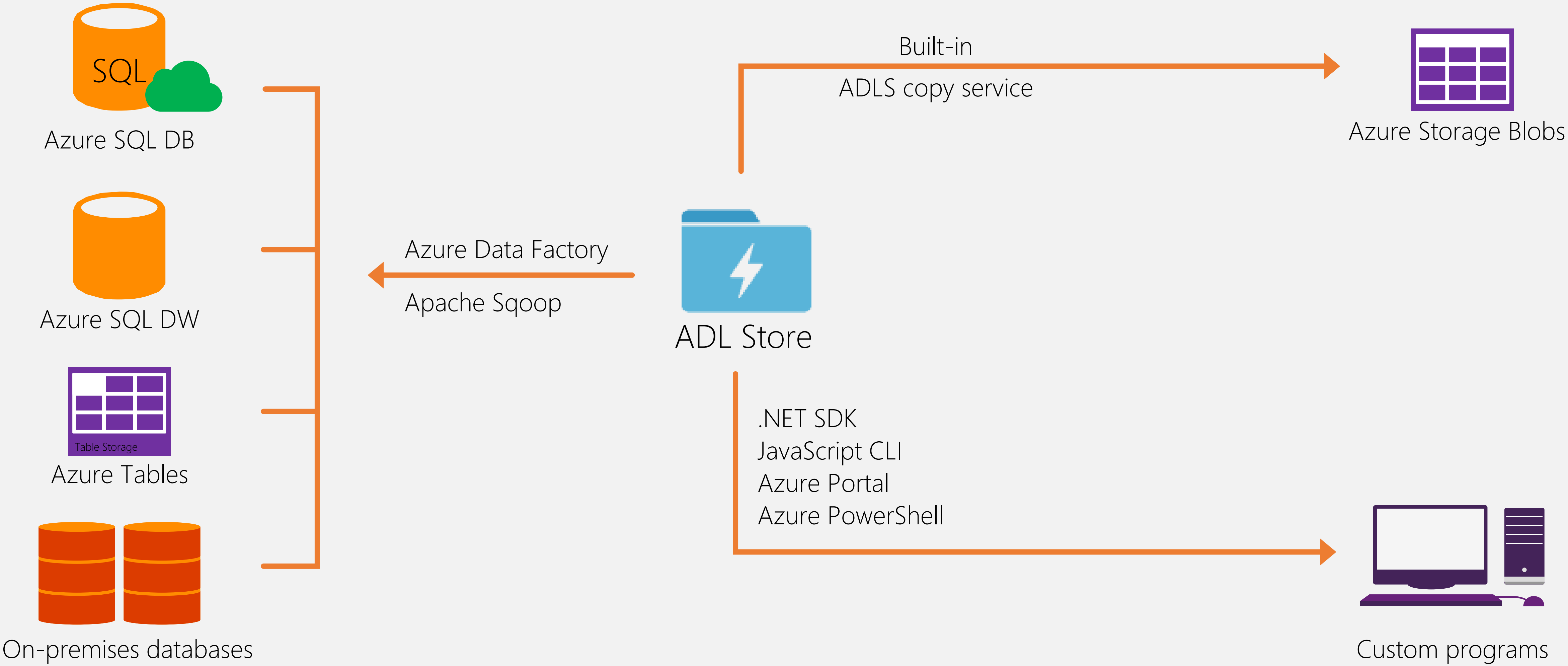
ADL Store: Ingress

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

Azure Data Lake Analytics

Microsoft

Azure Data Lake Analytics

Massively parallel, extensible, analytics made simple

The first cloud analytics service where you can easily develop and run massively parallel data transformation and processing programs in U-SQL, R, Python and .Net over petabytes of data. With no infrastructure to manage, process data on demand, scale instantly, and only pay per job.

Start in seconds, Scale instantly, Pay per job

Develop massively parallel programs with simplicity

Debug and Optimize your Big Data programs with ease

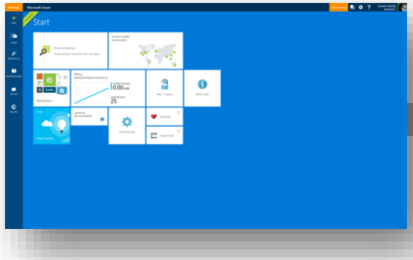
Virtualize your analytics

Enterprise-grade Support and Security

Start in seconds, Scale instantly, Pay per job

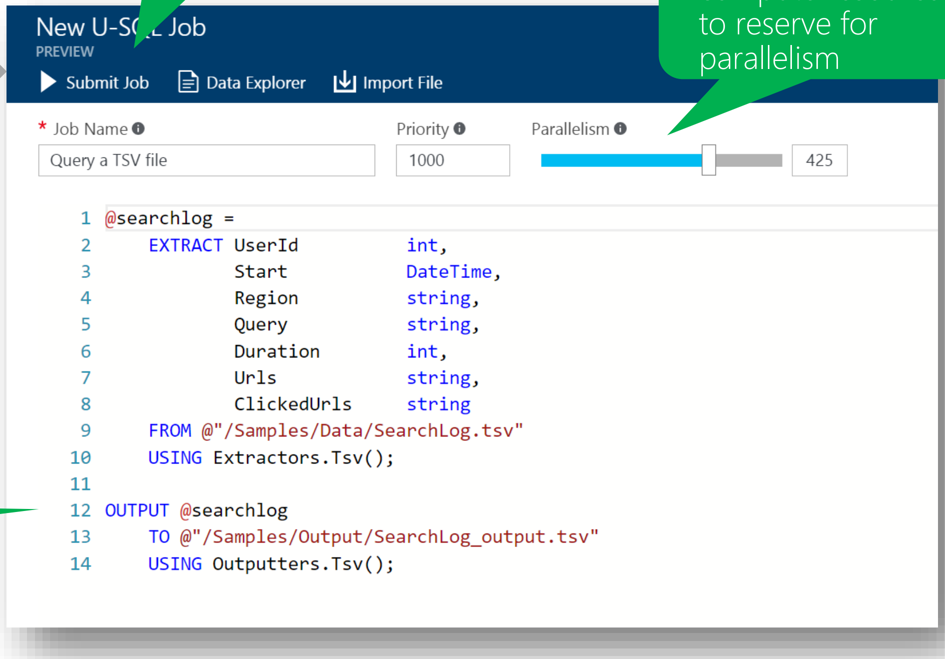
Our on-demand service will have you processing Big Data jobs **within 30 seconds**. There is no infrastructure to worry about because there are no servers, VMs, or clusters to wait for, manage or tune. You can **instantly scale the analytic units** (processing power) from one to thousands for each job. You only **pay for the processing used per job**.

1. In Azure Portal create a Data Lake Analytics Account



4. Submit the job

3. Choose how much compute resources to reserve for parallelism



* Job Name: Query a TSV file

Priority: 1000

Parallelism: 425

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

OUTPUT @searchlog
    TO @"/Samples/Output/SearchLog_output.tsv"
    USING Outputters.Tsv();
```

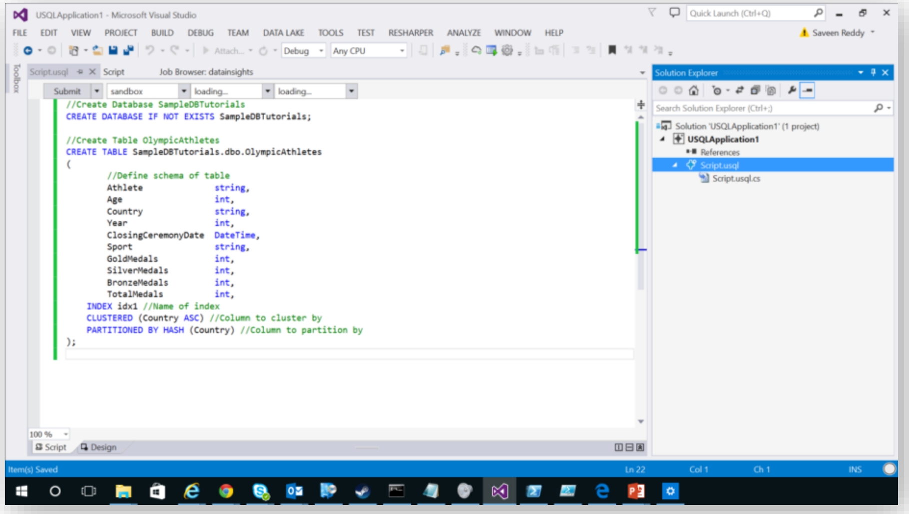
2. Write a big data program with U-SQL

DEMO

Microsoft

Develop massively parallel programs with simplicity

U-SQL is a simple, expressive, and extensible language that allows you to write code once and automatically have it be parallelized for the scale you need. You can process petabytes of data for diverse workload categories such as ETL, machine learning, cognitive science, machine translation, imaging processing, and sentiment analysis by using U-SQL and leveraging existing libraries written in .NET languages, R, or Python..

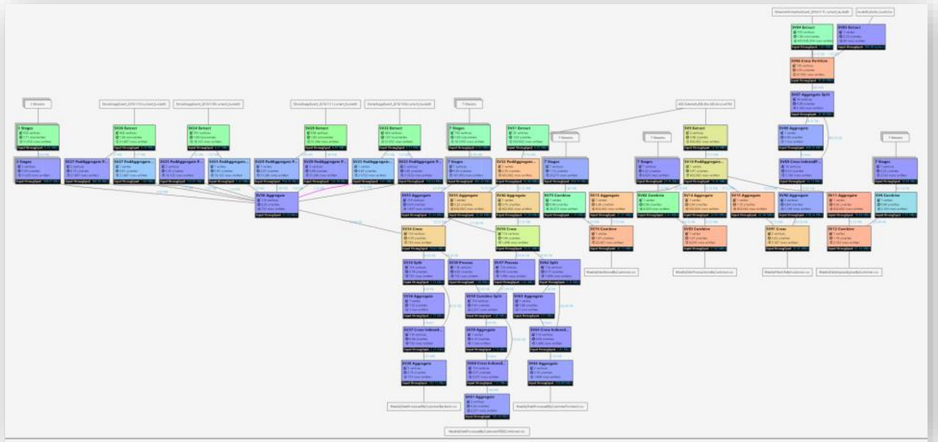


```
//Create Database SampleDBTutorials
CREATE DATABASE IF NOT EXISTS SampleDBTutorials;

//Create Table OlympicAthletes
CREATE TABLE SampleDBTutorials.dbo.OlympicAthletes
(
        //Define schema of table
        Athlete             string,
        Age                 int,
        Country             string,
        Year                int,
        ClosingCeremonyDate DateTime,
        Sport               string,
        GoldMedals          int,
        SilverMedals        int,
        BronzeMedals        int,
        TotalMedals         int,
    INDEX idx1 //Name of index
    CLUSTERED (Country ASC) //Column to cluster by
    PARTITIONED BY HASH (Country) //Column to partition by
);
```

Debug and Optimize your Big Data programs with ease

Debugging failures in cloud distributed programs are now as easy as debugging a program in your personal environment. Our execution environment actively analyzes your programs as they run and offers recommendations to improve performance and reduce cost. For example, if you requested 1000 AUs for your program and only 50 AUs were needed, the system would recommend that you only use 50 AUs resulting in a 20x cost savings.

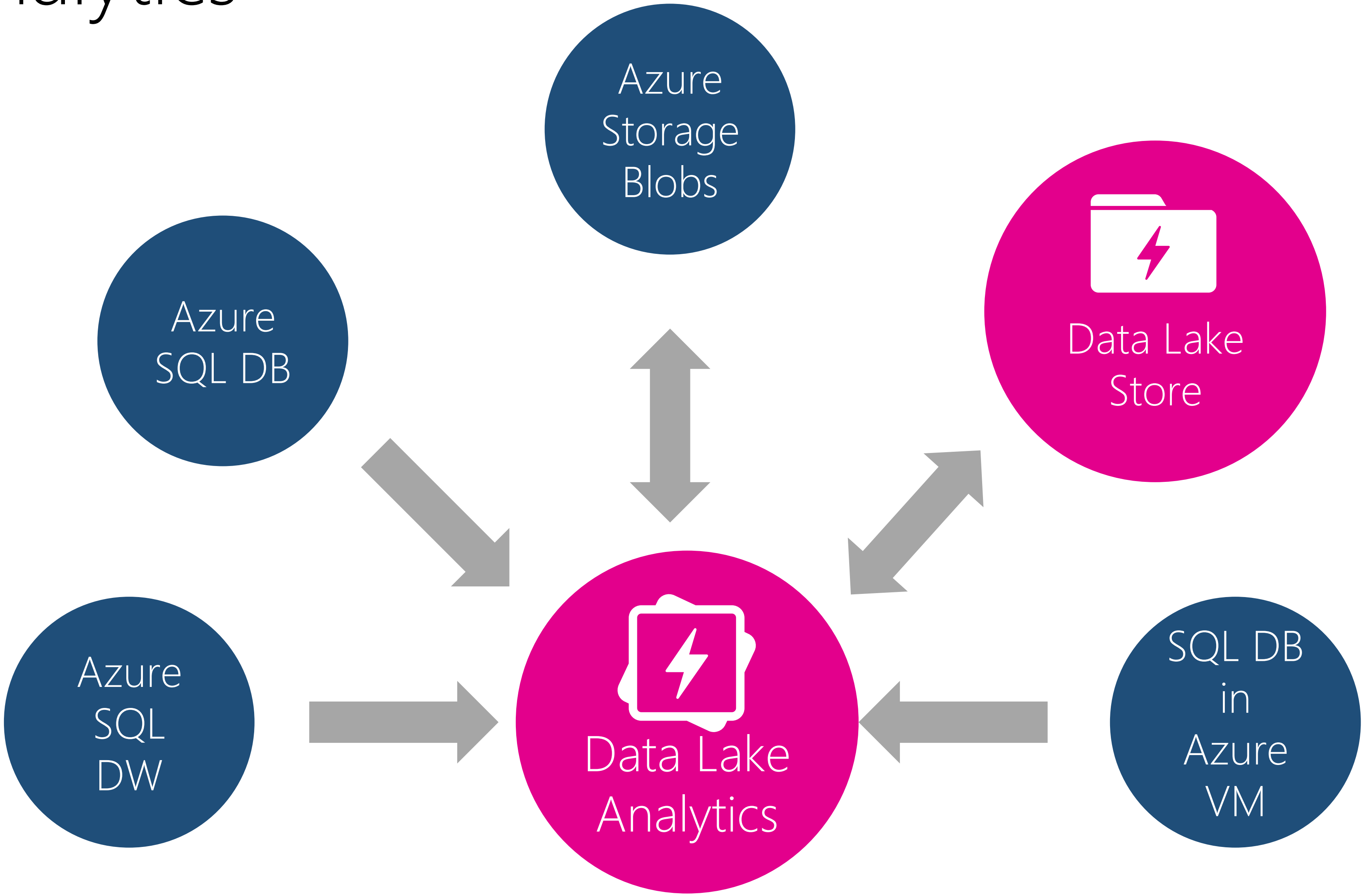


DEMO

Microsoft

Virtualize your analytics

U-SQL can use data from sources in Azure. Where possible data transformation is pushed close to the source data to minimize data transfer and maximize performance.



Microsoft

U-SQL

A new language for Big Data

Familiar syntax to millions of SQL & .NET developers

Unifies declarative nature of SQL with the imperative power of C#

Unifies structured, semi-structured and unstructured data

Distributed query support over all data

Microsoft

Language Overview

U-SQL Fundamentals

- All the familiar SQL clauses

  SELECT | FROM | WHERE
  GROUP BY | JOIN | OVER
- Operate on unstructured and structured data
- Relational metadata objects

.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 | I/O (Extractors, Outputters)

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

U-SQL Example

```
DECLARE @endDate DateTime = DateTime.Now;
DECLARE @startDate DateTime = @endDate.AddDays(-7);


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


@orders = SELECT * FROM orders
     WHERE Date >= startDate AND Date <= endDate;


@orders = SELECT * FROM orders
     WHERE Customer.Contains(“Contoso”);


OUTPUT @orders
    TO "/output/output.txt"
    USING Outputters.Tsv();
```

DECLARE constant values

C# Expressions

RowSets enable dataflow programming

EXTRACT performs schema on read for files

OUTPUT for writing files

Built-in handling for CSV & TSV

Whole-script optimization

U-SQL Example

```
CREATE ASSEMBLY OrdersDB.SampleDotNetCode
    FROM @"/DLLs/Helpers.dll";


REFERENCE ASSEMBLY OrdersDB.Helpers;


@rows =
    SELECT
        OrdersDB.Helpers.Normalize(Customer) AS Customer,
        Amount AS Amount
    FROM @orders;


@rows =
    PROCESS @rows
    PRODUCE OrderId string, FraudDetection double
    USING new OrdersDB.Detection.FraudAnalyzer();

OUTPUT @rows
    TO "/output/output.dat"
    USING OrdersDB.CustomOutputter();
```

CREATE ASSEMBLY to register code in the U-SQL Catalog

REFERENCE ASSEMBLY to bring code to each container (vertex)

Directly call C# methods from Assemblies in U-SQL Expressions

PROCESS to perform row-by-row transformation of data with User-Defined Operator UDO

OUTPUT with custom Outputter UDO to support user-defined formats.

U-SQL Analytics

Traditional ETL/Analytics

Aggregation/Grouping

Window Functions: Ranking, percentiles, etc.

Massively parallelized execution of …

- .NET
- R
- Python,
- etc.

U-SQL/Cognitive

Built-in Cognitive capabilities in U-SQL utilizing the same algorithms powering Cognitive services.

- Vision (Face, Emotion, Tagging, OCR)
- Text Analytics (Key Phrase Extraction, Sentiment Analysis)

Fundamentals

Microsoft

Usage-based Billing

HDInsight

Billed for

Cluster time

Data Lake Analytics

Billed for

Compute resources used for the duration of your query

Data Lake Store

Billed for

- Amount of data stored
- Number of I/O operations

Microsoft

Enterprise-Grade Support & Security

99.9% Uptime

24/7 Support

Defense-in-depth

- Perimeter security
- Azure AD Auth
- RBAC & ACLs
- Encryption at rest and on the wire
- All operations Audited

ADL Analytics Compared To….

HDInsight

clusters as service

- Customization
- Control
- Flexibility
- Leverage the Hadoop ecosystem (Spark, Storm, Hbase, …)

ADL Analytics

big data queries as a service

- Convenience
- Efficiency
- Automatically scale
- A simple way to get started using familiar concepts, languages, and tools

Mix and match them depending on what you need.

Share data in the Azure Data Lake Store

Azure SQL DW

- SQL Language & Tools
- Scale DW with DWUs
- Massively parallelize SQL

ADL Analytics

- U-SQL (+Hive in the future) & Open Source Tools
- Scale per query (1 to 1000+ nodes)
- Interoperate with Open Source technologies and data formats (Mahout, ORC, Parquet,…)
- Massively parallelize .NET code, R, Python

Mix and match them depending on what you need.

Share data in the Azure Data Lake Store

Q & A

Microsoft

http://aka.ms/AzureDataLake

Microsoft

Microsoft

Backup