Moving data between ADLS and other systems

Mithun Prasad, PhD
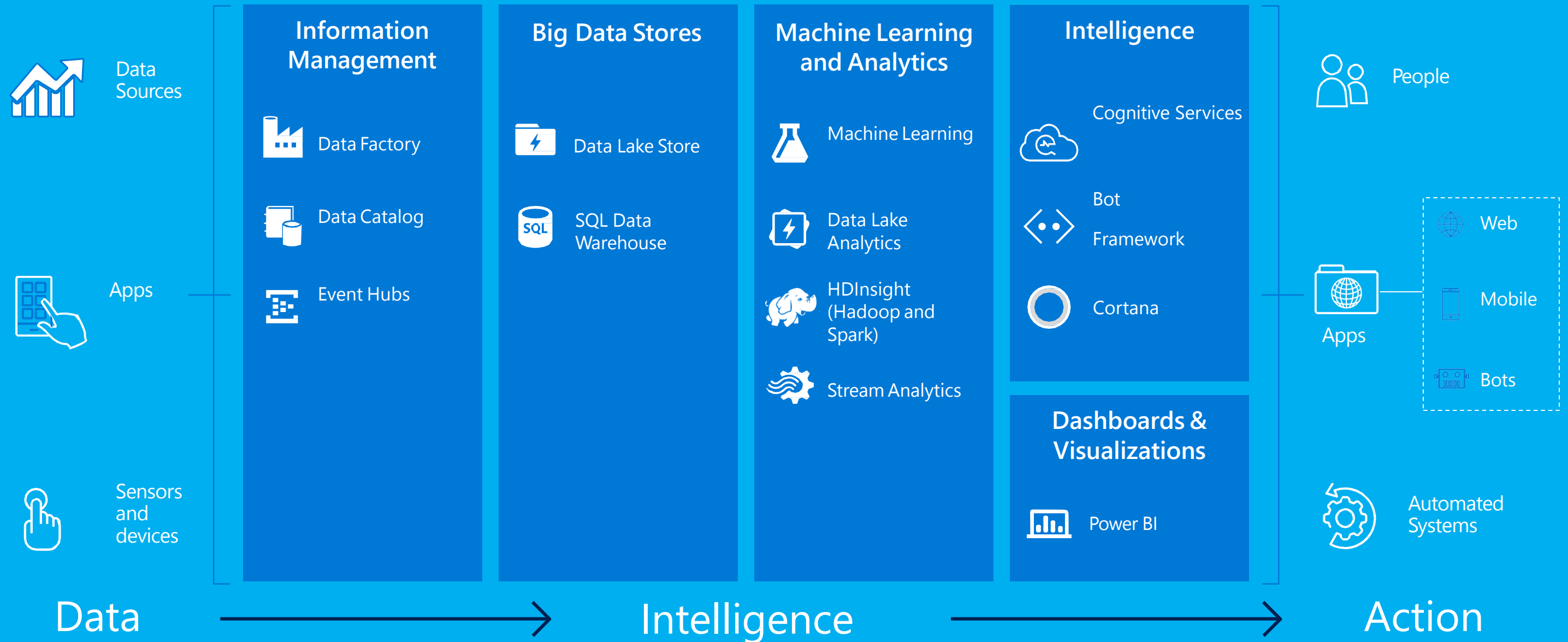
Senior Program Manager @ Microsoft

Agenda

- Overview
- Azure Data Lake Store PowerShell cmdlets
- AdlCopy
- Sqoop
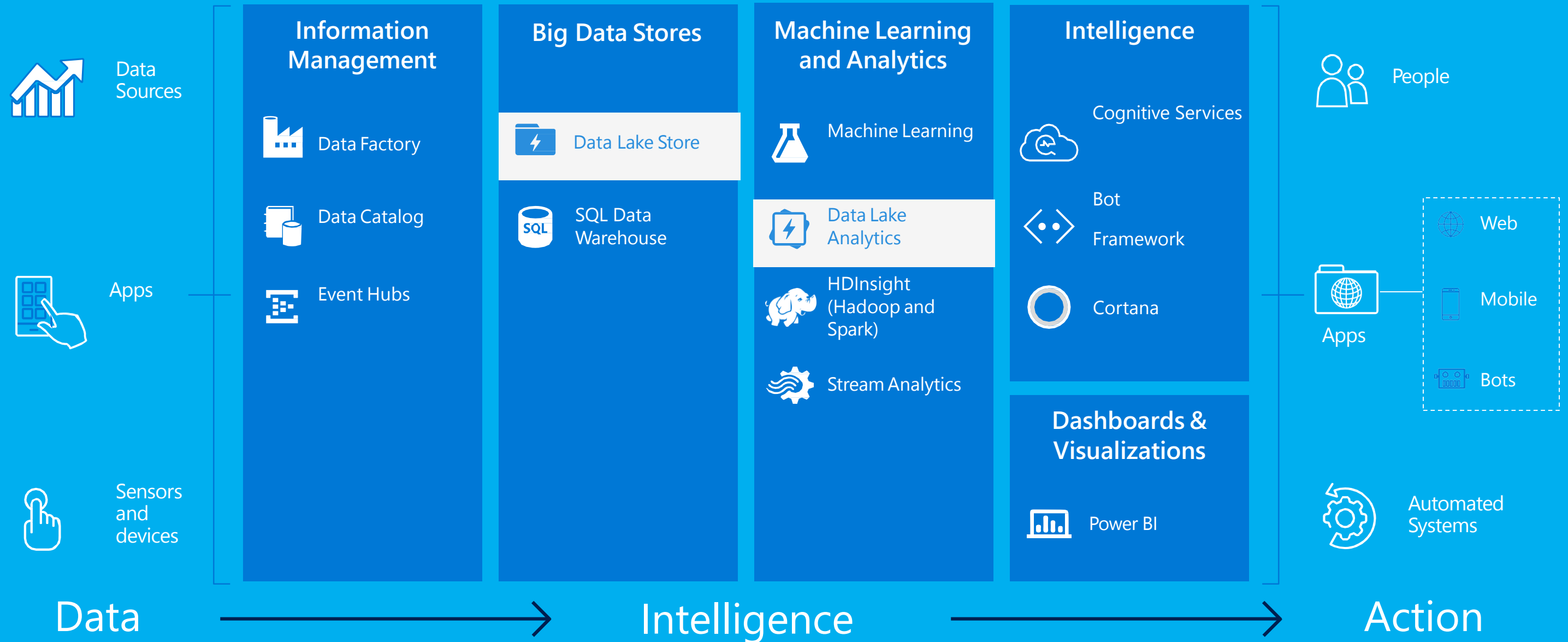- DistCP
- Azure Data Factory (ADF)

Cortana Intelligence Suite
Transform data into intelligent action

Data Sources

Apps

Sensors and devices

Information Management

- Data Factory
- Data Catalog
- Event Hubs

Big Data Stores

- Data Lake Store
- SQL Data Warehouse

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

Cortana Intelligence Suite
Transform data into intelligent action

Data Sources

Apps

Sensors and devices

Information Management

- Data Factory
- Data Catalog
- Event Hubs

Big Data Stores

- Data Lake Store
- SQL Data Warehouse

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

Web

Mobile

Bots

Apps

Automated Systems

Data ⟶ Intelligence ⟶ Action

Big Data Analytics – Data Flow

Data sources

Apps

Sensors and devices

Ingest data

Process data

Download data

Visualize data
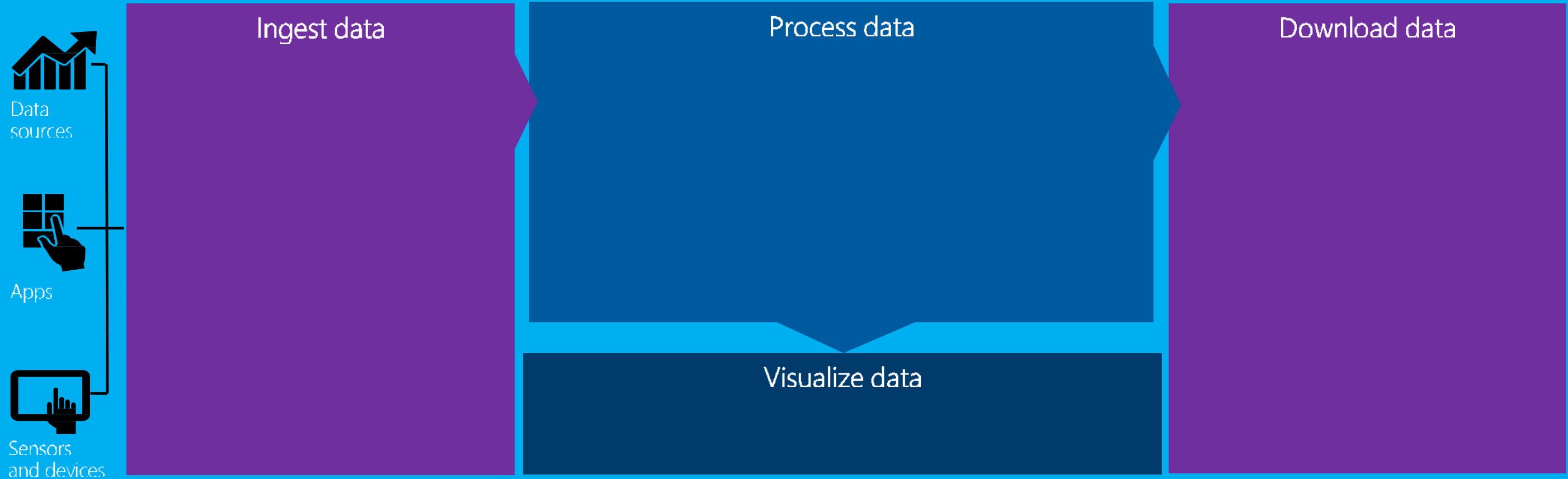
Ingestion

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

Azure SQL DB

Azure SQL DW

Azure tables

On-premises databases

Azure Data Factory

ADL Store

Azure Stream Analytics

Azure Event Hubs

ADL built-in copy service
Azure Data Factory
Hadoop DistCp

Azure   Blobs

.NET SDK
CLI
Azure Portal
Azure PowerShell

Custom programs
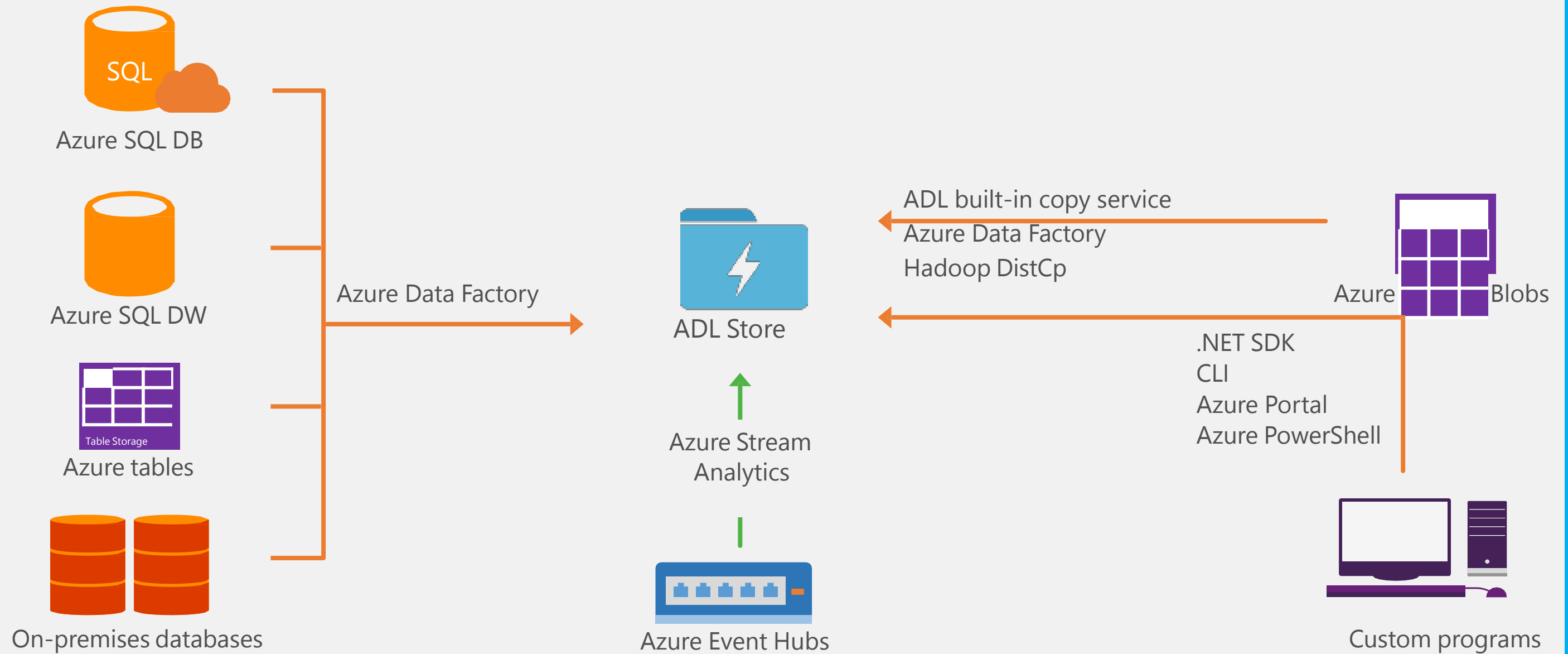
ADLS Ingestion – Getting started

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
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	Intuitive CopyData wizard for adhoc copies
AdlCopy	Copy data easily from Azure Storage at least cost
Distcp on HDI	Want to analyze data using HDI Familiarity with OSS tools

ADLS Ingestion – Orchestration

Out-of-the-box tools

Azure Data Factory
First-class support for ADLS
Support variety of endpoints
    WASB, OnPrem, Relational DB
Integrated with Analytic tools
Programmatic customization

OSS tools
Sqoop - Copy from relational DBs
Distcp - Copy from WASB
Use Oozie & Falcon on HDI to manage
Use Storm for streaming data from Eventhub / Kafka into ADLS

PowerShell
Use built-in commandlets
Use PowerShell Workflow Runbooks to manage
Use PowerShell Script Runbooks to manage

Azure Import/Export Service
Transfer data using disks if ExpressRoute is not feasible
Lots of manual steps and not easy to orchestrate

Custom & LOB Apps

ADLS SDK
Available in various languages (.NET, Java, Node.js, ..)
Upload from distributed sources e.g. server logs
ADF can be used to manage .NET apps

REST APIs
For unsupported languages and platforms
Will need custom apps for orchestration

Azure Data Lake Store PS Cmdlets

Export-AzureRmDataLakeStoreItem   Downloads a file from Data Lake Store.

Get-AzureRmDataLakeStoreItemContent   Gets the contents of a file in Data Lake Store.

Import-AzureRmDataLakeStoreItem   Uploads a file to Data Lake Store.

Join-AzureRmDataLakeStoreItem   Joins one or more files to create one file in Data Lake Store.

Move-AzureRmDataLakeStoreItem   Moves or renames a file or folder in Data Lake Store.

New-AzureRmDataLakeStoreItem   Creates a new file or folder in Data Lake Store.

Remove-AzureRmDataLakeStoreItem   Deletes a file or folder in Data Lake Store.

More details are here

AdlCopy

- More details are [here](#)
- Primarily it is command line tool, [AdlCopy](#), to copy data very easily from Azure Storage Blobs / Data Lake Store into Data Lake Store.
- You can use the AdlCopy tool in two ways:
  - **Standalone**, where the tool uses Data Lake Store resources to perform the task.
  - **Using a Data Lake Analytics account**, where the units assigned to your Data Lake Analytics account are used to perform the copy operation. You might want to use this option when you are looking to perform the copy tasks in a predictable manner.

```
AdlCopy /Source <Blob source> /Dest <ADLS destination>
/SourceKey <Key for Blob account> /Account <ADLA account>
/Units <Number of Analytics units>
```

DistCP

- You can use DistCP to copy data **to and from** an HDInsight cluster storage (WASB) into a Data Lake Store account.
- More details are [here](#).

```
hadoop distcp
wasb://<container_name>@<storage_account_name>.blob.core.windows.net/example/data/gutenberg
adl://<data_lake_store_account>.azuredatalakestore.net:443/myfolder
```

Sqoop

- [Apache Sqoop](#) is a tool designed to transfer data between relational databases and a big data repository, such as Data Lake Store.

- You can use Sqoop to copy data **to and from** Azure SQL database into a Data Lake Store account, in addition to other other relational DBs.

- More details are [here](#).

```
sqoop-import --connect "jdbc:sqlserver://<sql-database-server-name>.database
.windows.net:1433;username=<username>@<sql-database-server-name>;password=
<password>;database=<sql-database-name>" --table Table1 --target-dir adl://
<data-lake-store-name>.azuredatalakestore.net/Sqoop/SqoopImportTable1
```
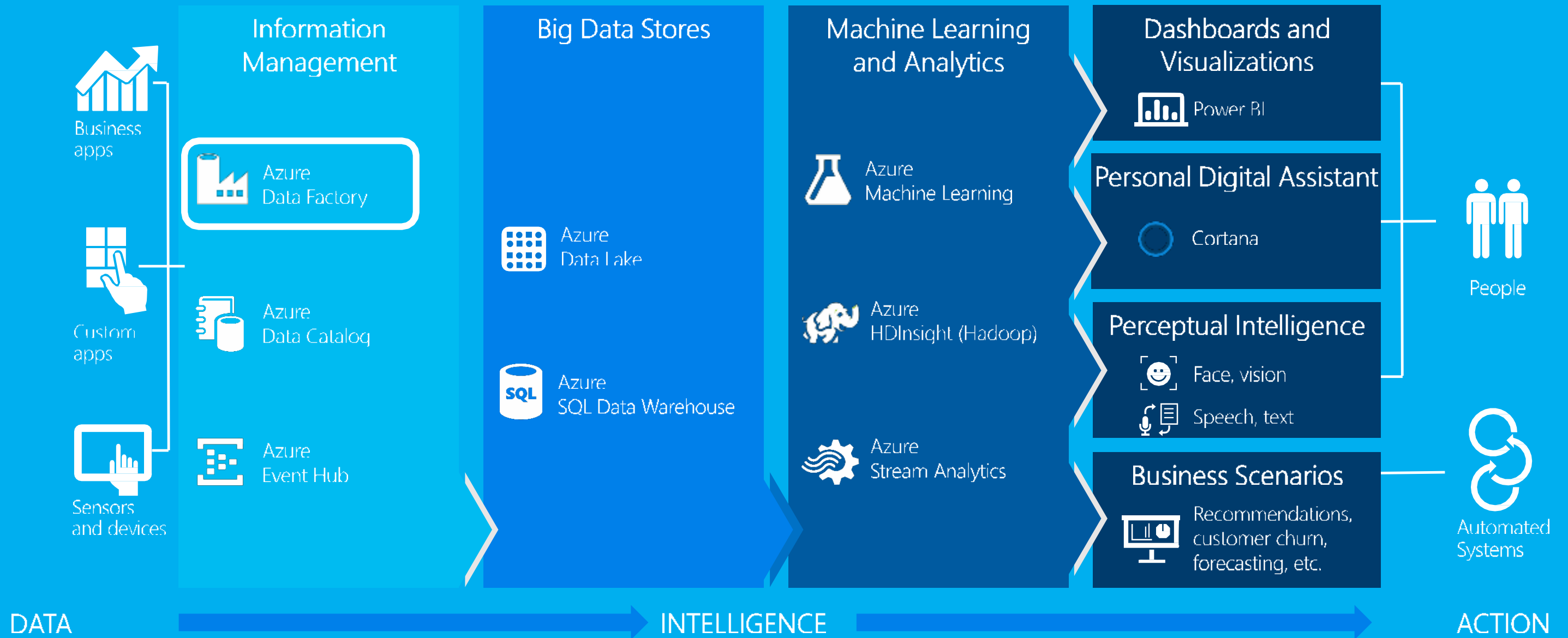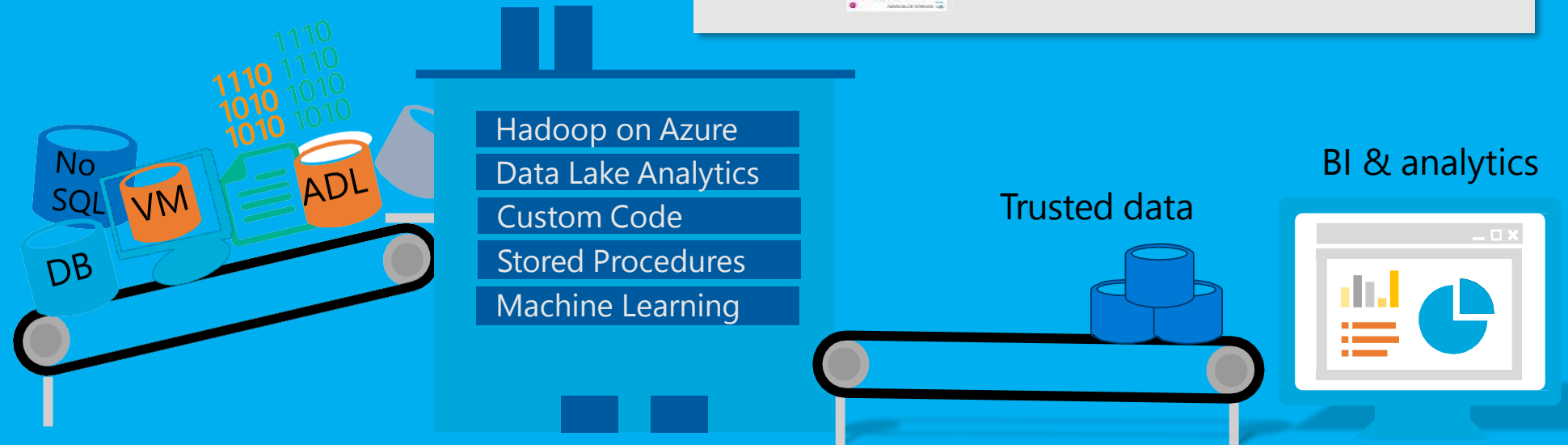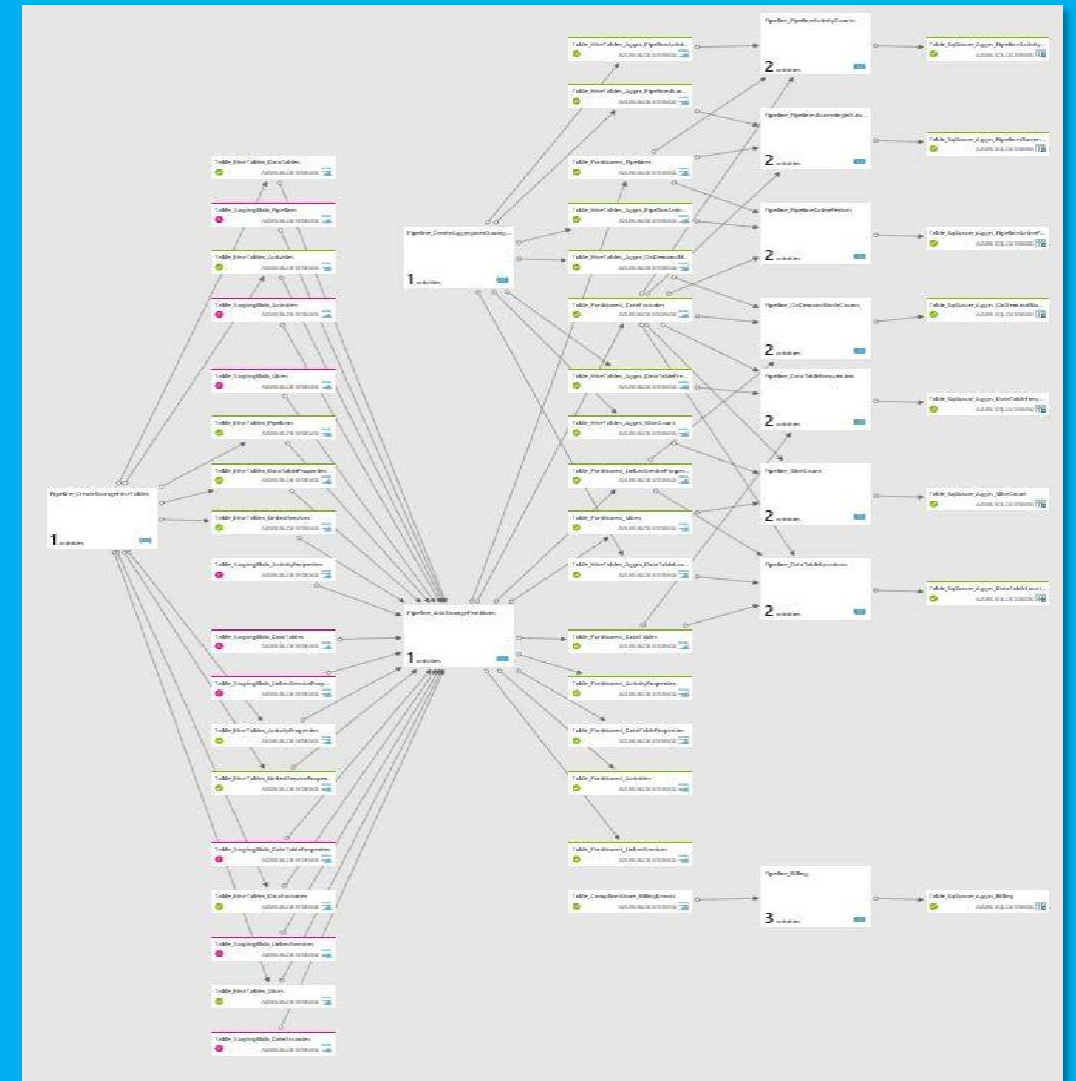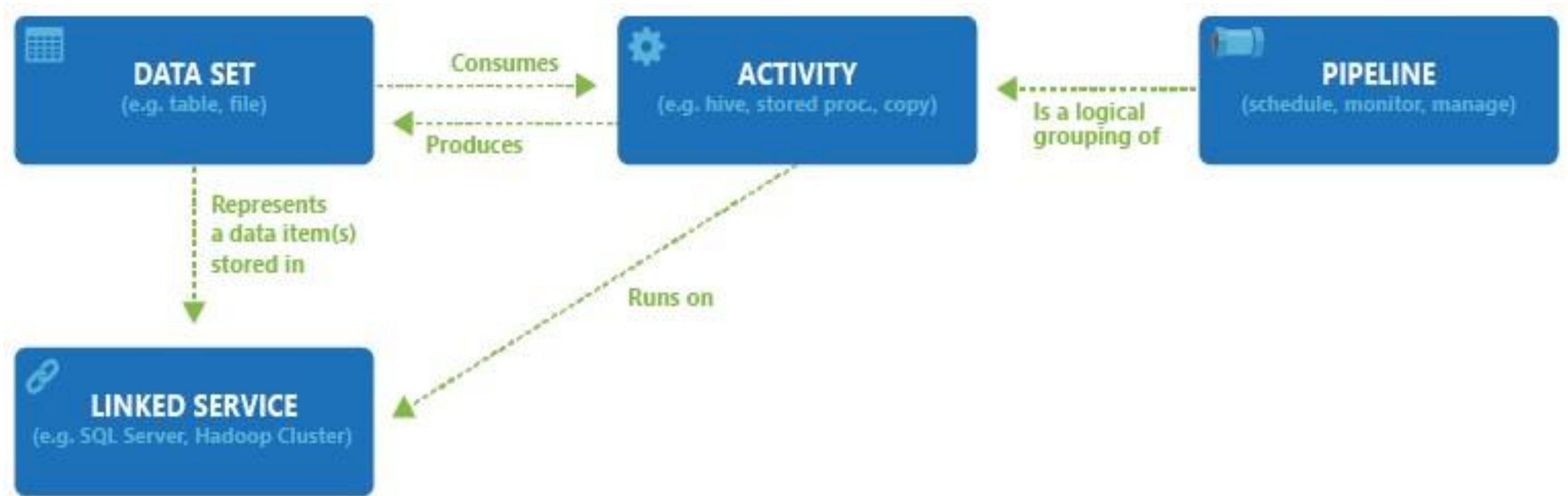
Azure Data Factory

Compose, orchestrate & monitor data services at scale

- Fully managed service to support orchestration of

  data movement and transformation

- Connect to relational or non-relational data that is on-

  premises or in the cloud

- Cost Effective

- SSIS vs Data Factory

1110
1110
1110
1110
1010 1010
1010 1010

No SQL
VM
ADL
DB

Hadoop on Azure
Data Lake Analytics
Custom Code
Stored Procedures
Machine Learning

Trusted data

BI & analytics

Azure Data Factory Components

DATA SET (e.g. table, file) —Consumes→ **ACTIVITY** (e.g. hive, stored proc., copy) ←Produces— **DATA SET**

ACTIVITY ←Is a logical grouping of— **PIPELINE** (schedule, monitor, manage)

DATA SET —Represents a data item(s) stored in→ **LINKED SERVICE** (e.g. SQL Server, Hadoop Cluster)

ACTIVITY —Runs on→ **LINKED SERVICE**

Microsoft

Azure Data Factory

Connects Azure 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*	✓	
	PostgreSQL*	✓	
	Sybase*	✓	
	Cassandra*	✓	
	MongoDB*	✓	
	Amazon Redshift	✓	
File	File System*	✓	✓
	HDFS*	✓	
	Amazon S3	✓	
Others	Salesforce	✓	
	Generic ODBC*	✓	
	Generic OData	✓	
	Web Table (table from HTML)	✓	
	GE Historian*	✓	

* Can be on-premises or on Azure IaaS, enabled using Data Management Gateway

Azure Data Factory Customers

ADF Resources

Get Started
- [Learning Path](#)
- [Azure Portal](#) & Customer Profiling
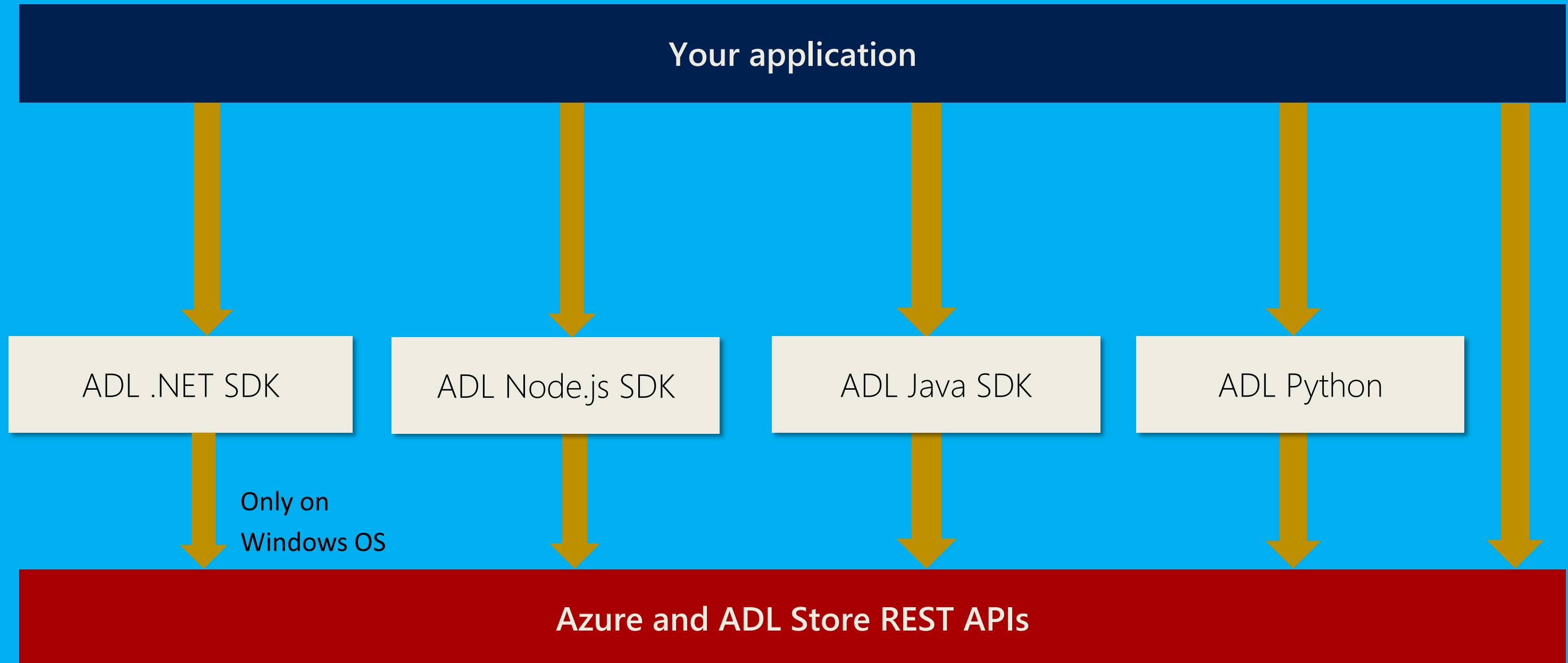- [Visual Studio plug-in installation](#)

Learn More
- [Product Recommendations Virtual Lab MS Technet](#)
- [Azure Big Data Blog](#)

Give Feedback
- [UserVoice](#)

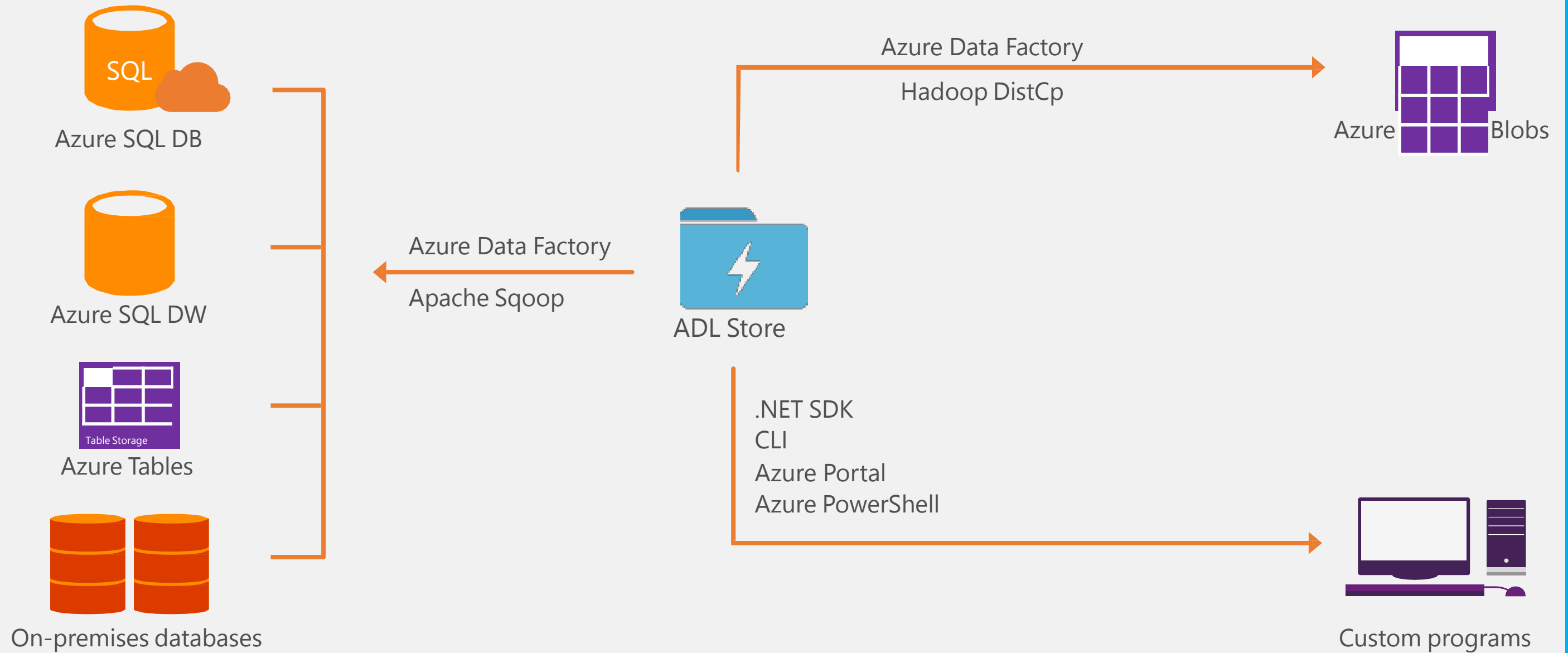## Get Help
- [ADF Forums](#)

Customizing using SDKs/APIs

Your application

| ADL .NET SDK | ADL Node.js SDK | ADL Java SDK | ADL Python |

Only on

Windows OS

Azure and ADL Store REST APIs

Egress

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

SQL
Azure SQL DB

Azure SQL DW

Table Storage
Azure Tables

On-premises databases

Azure Data Factory

Hadoop DistCp

Azure Data Factory

Apache Sqoop

ADL Store

.NET SDK
CLI
Azure Portal
Azure PowerShell

Azure Blobs

Custom programs

Big Data Analytics – Data Flow

Data sources

Apps

Sensors and devices

Ingest data

Bulk Ingestion

Azure Data Factory

Developer SDKs

Portal & Tools

Apache DistCp, Sqoop

AdlCopy Service

Import/Export Service

Event Ingestion

STORM

Stream Analytics

Process data

Azure Data Lake Analytics

Azure HDInsight

Hadoop

HBASE

STORM

Azure Data Lake Store

Visualize data

Azure SQL DW

+

Power BI

Download data

Azure Data Factory

Developer SDKs

Portal & Tools

Apache DistCp, Sqoop

Import/Export Service

In conclusion

- Many tools are available to copy data into ADLS and out of ADLS
  - More information is here
  - You can use generic Azure tools e.g. PowerShell
  - You can use Open Source tools e.g. DistCP
  - You can use special purpose tools e.g. ADF
  - You can create your own tools
- Pick the tool that meets your scenario requirements
  - Optimize it on the dimension you want – Cost, Performance. Etc.

Questions?

http://aka.ms/AzureDataLake

Microsoft