

# Moving data between ADLS and other systems

Mithun Prasad, PhD

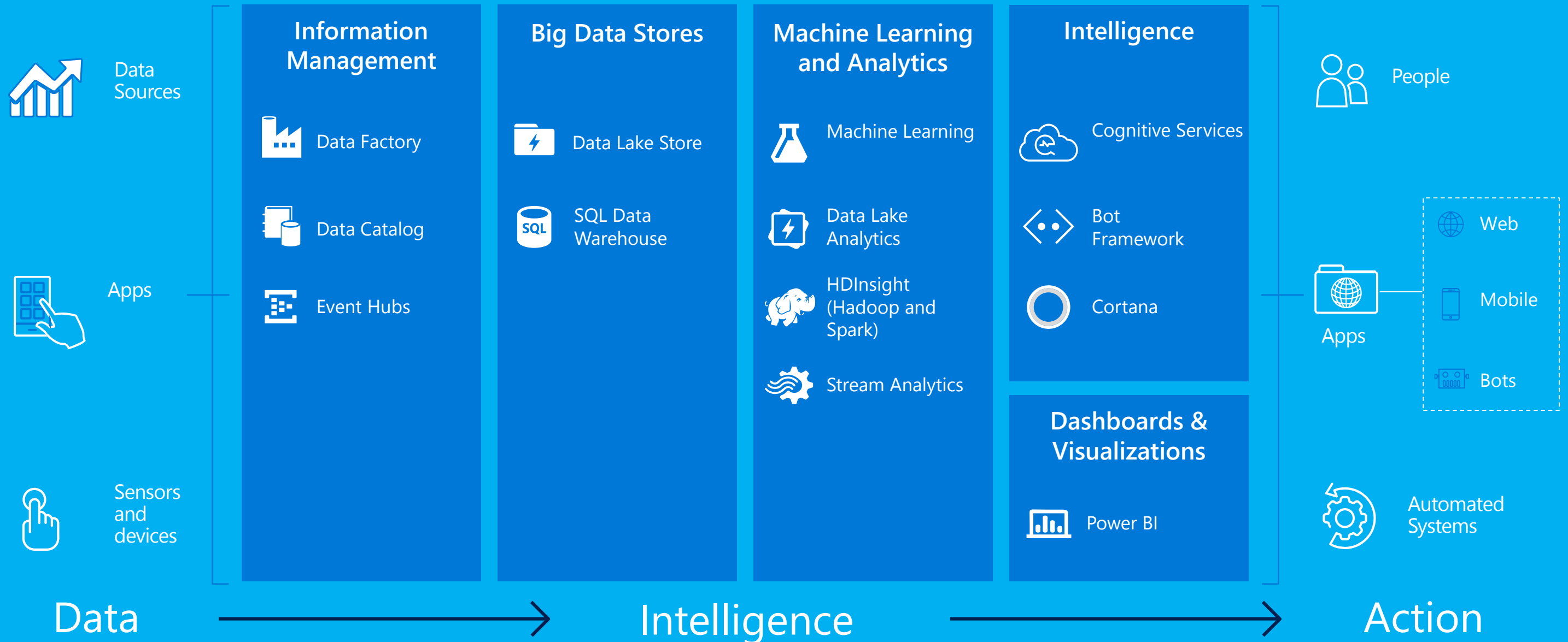
Senior Program Manager @ Microsoft

# Agenda

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

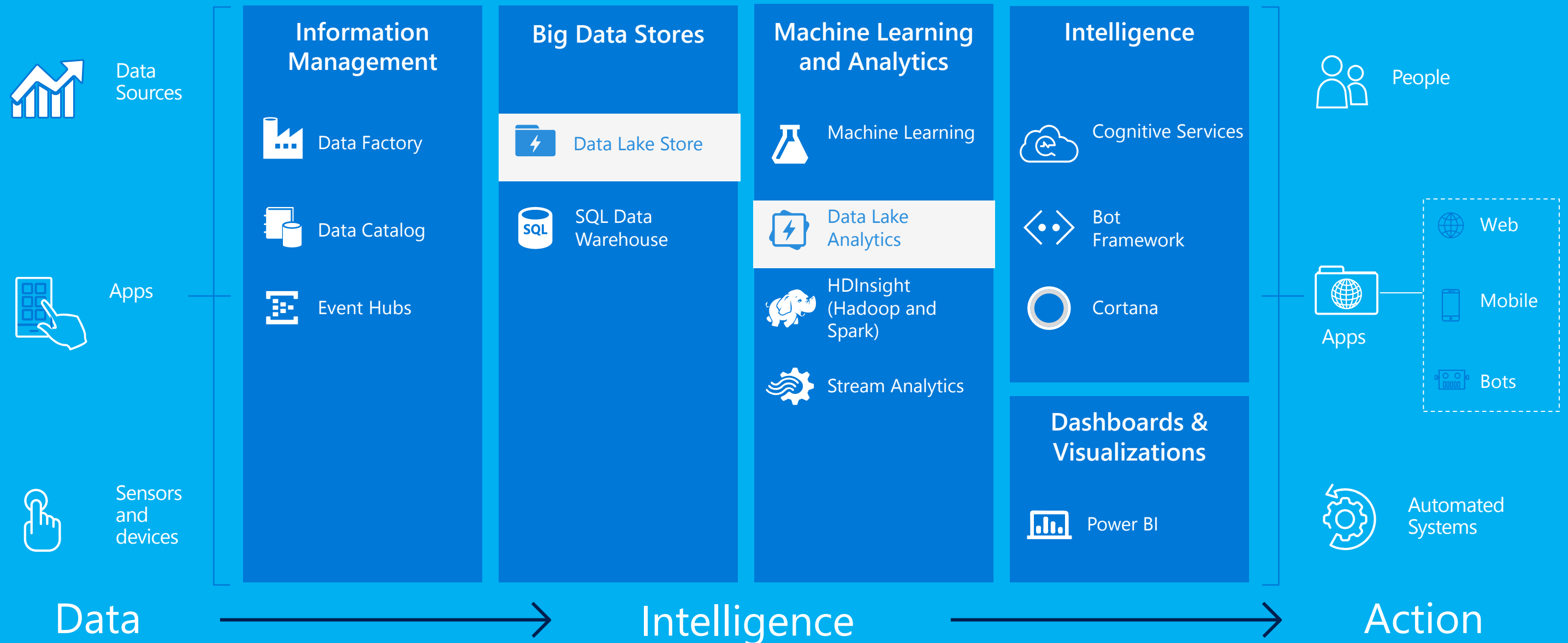
# Cortana Intelligence Suite

## Transform data into intelligent action

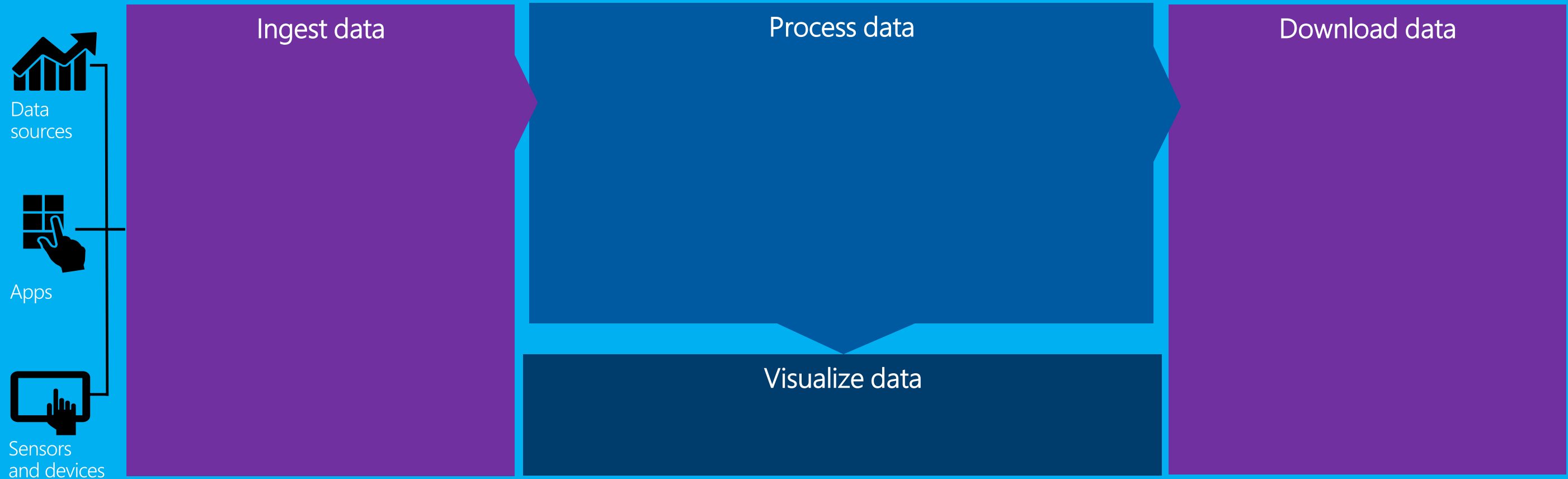


# Cortana Intelligence Suite

## Transform data into intelligent action

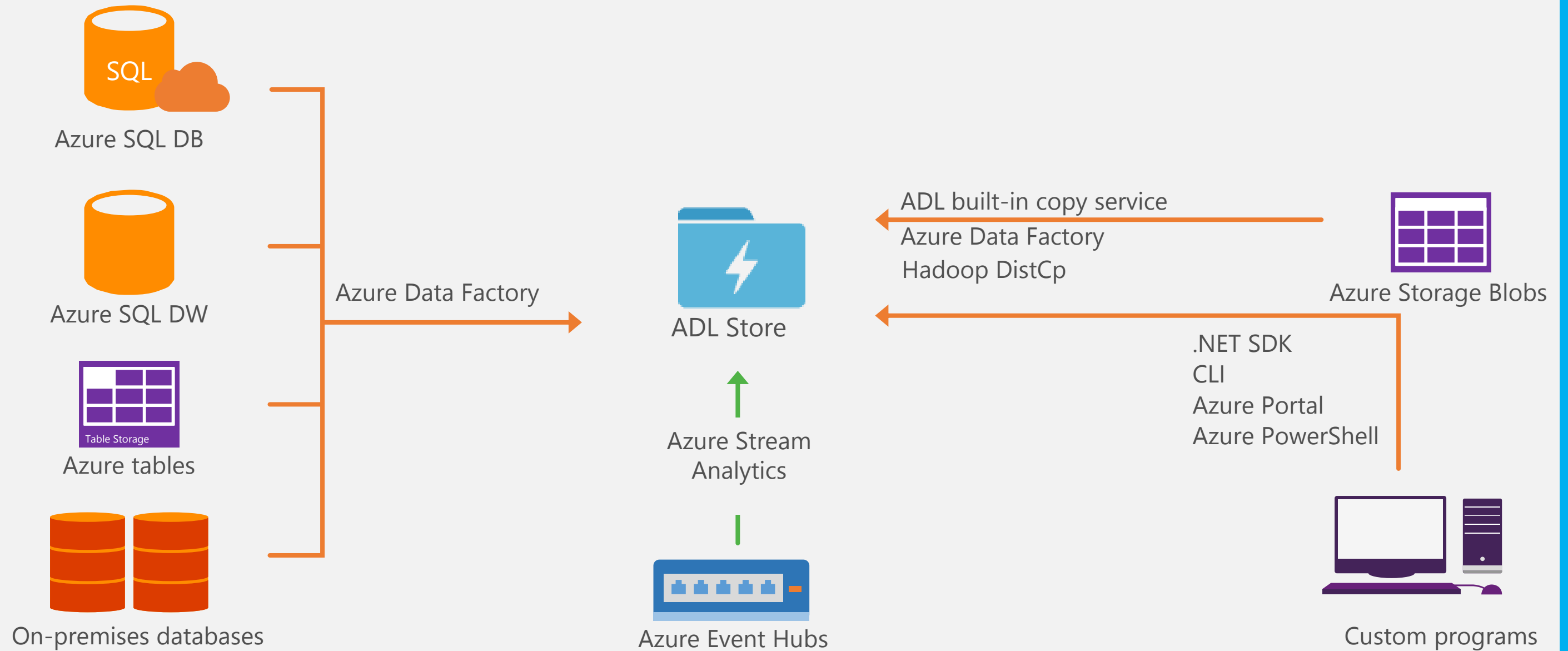


# Big Data Analytics – Data Flow

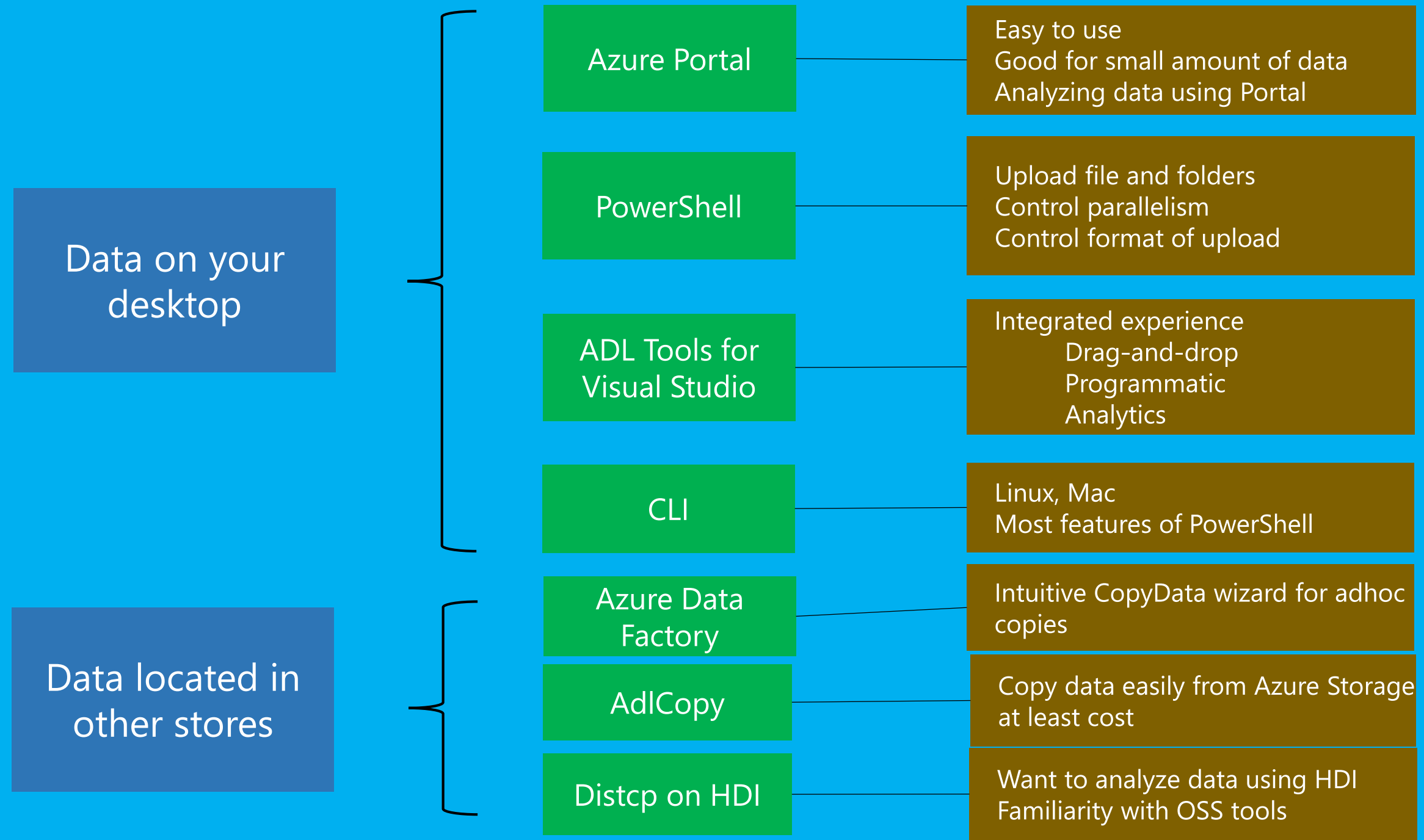


# Ingestion

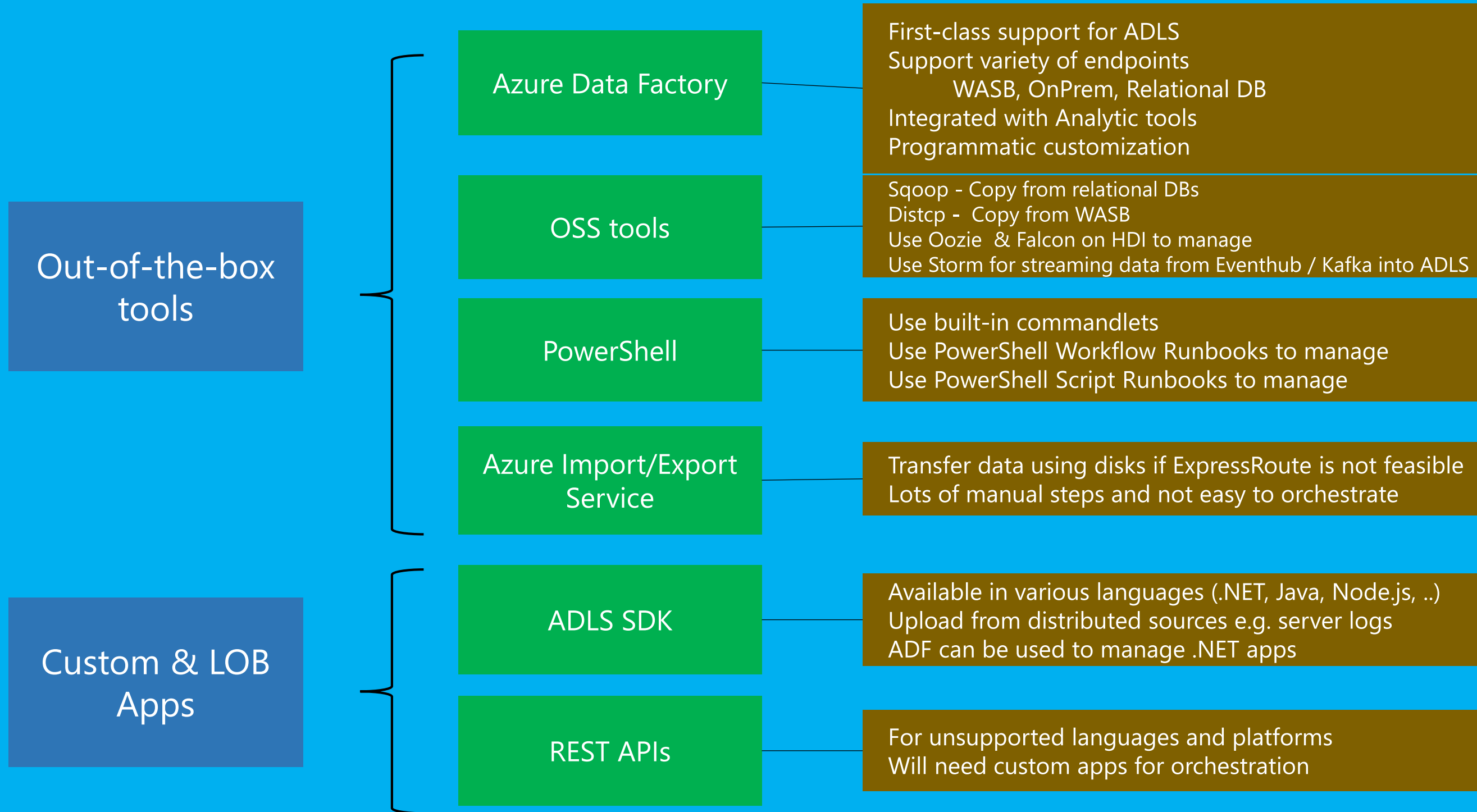
Data can be ingested into Azure Data Lake Store from a variety of sources



# ADLS Ingestion – Getting started



# ADLS Ingestion – 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
```

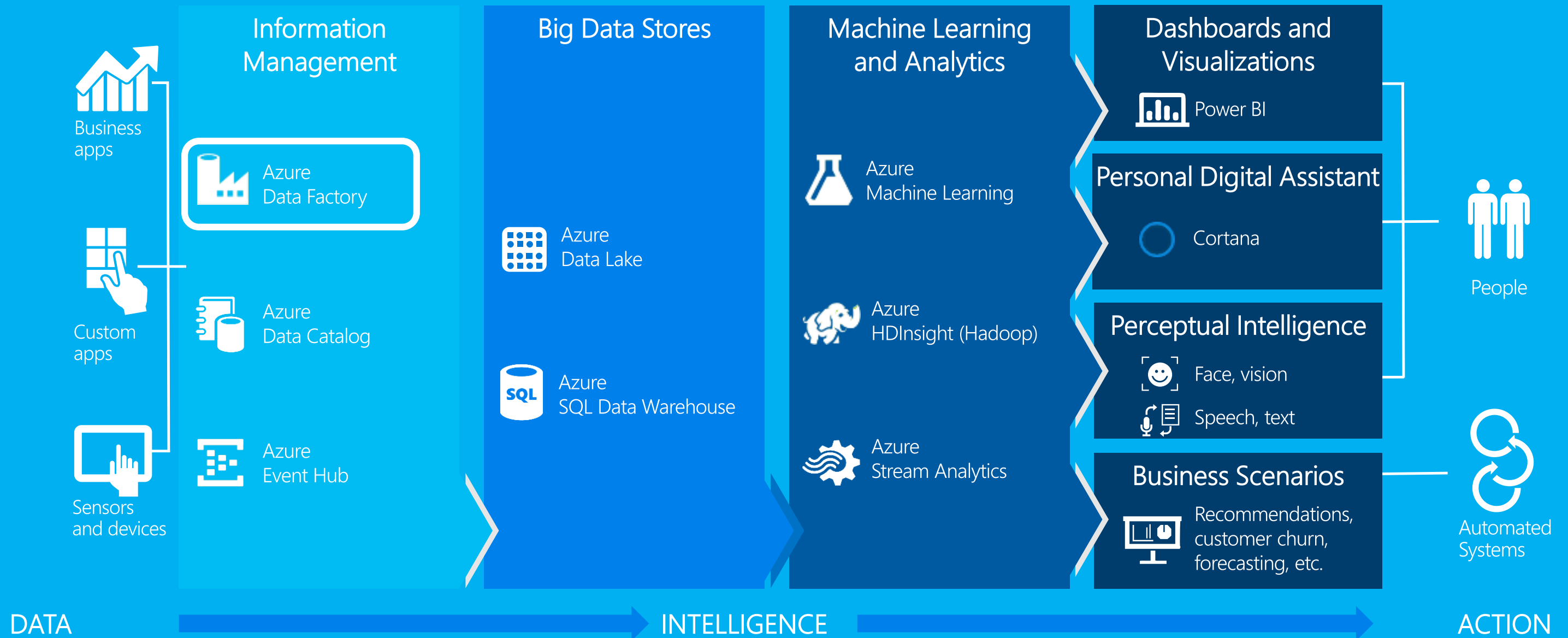
# Sqoop

- [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
```

# Cortana Intelligence Suite

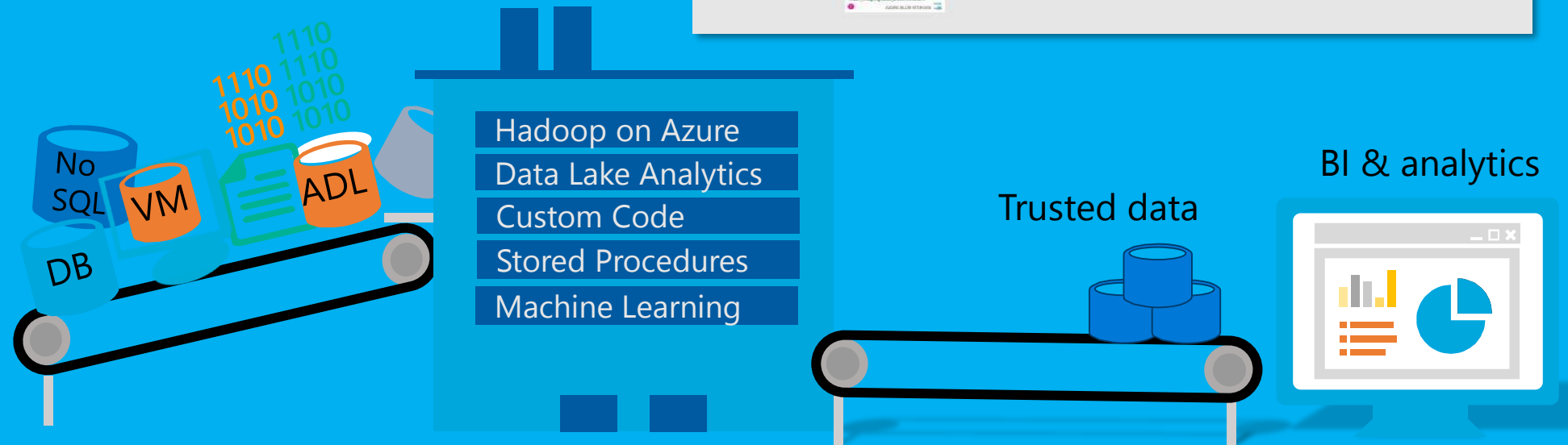
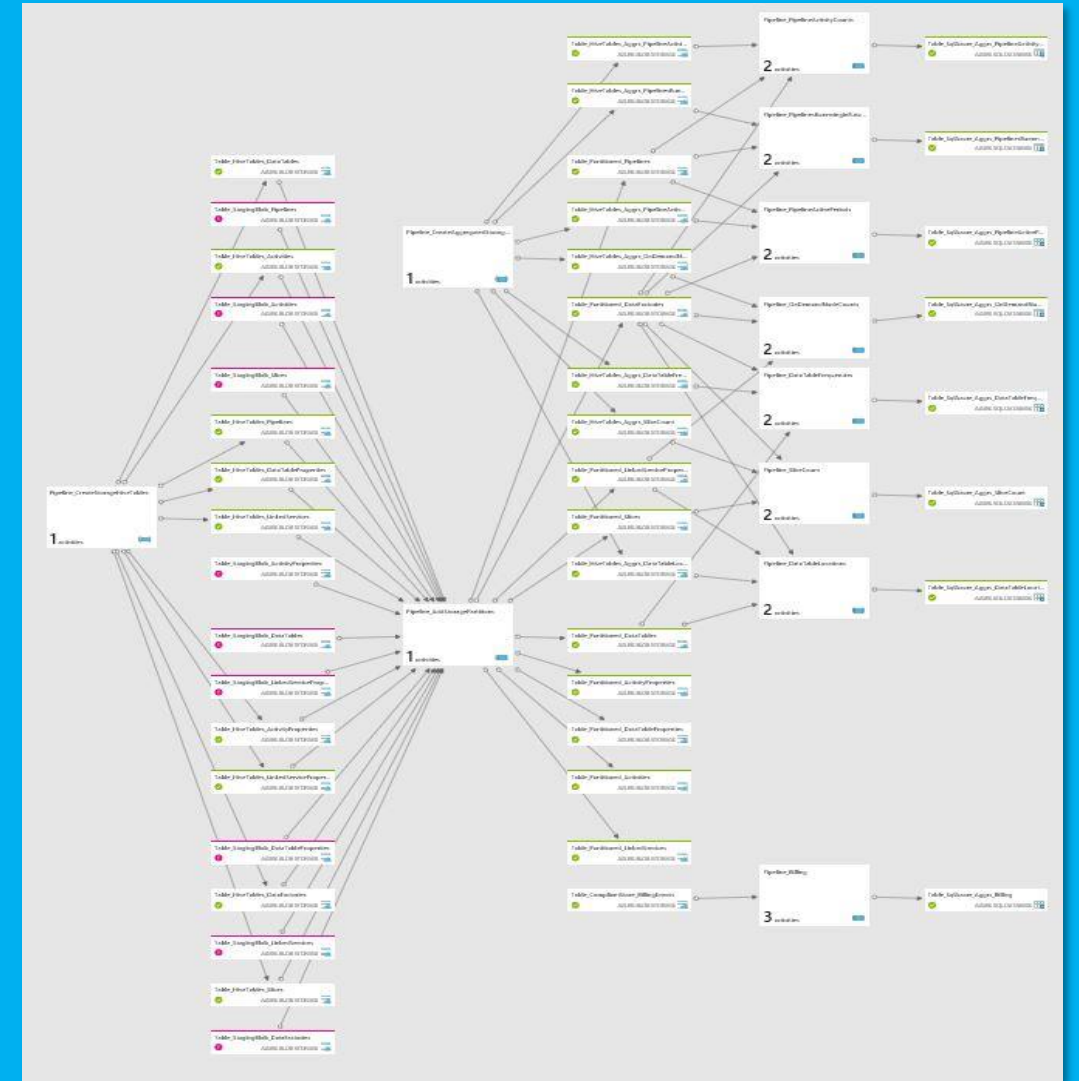
## Transform data into intelligent action



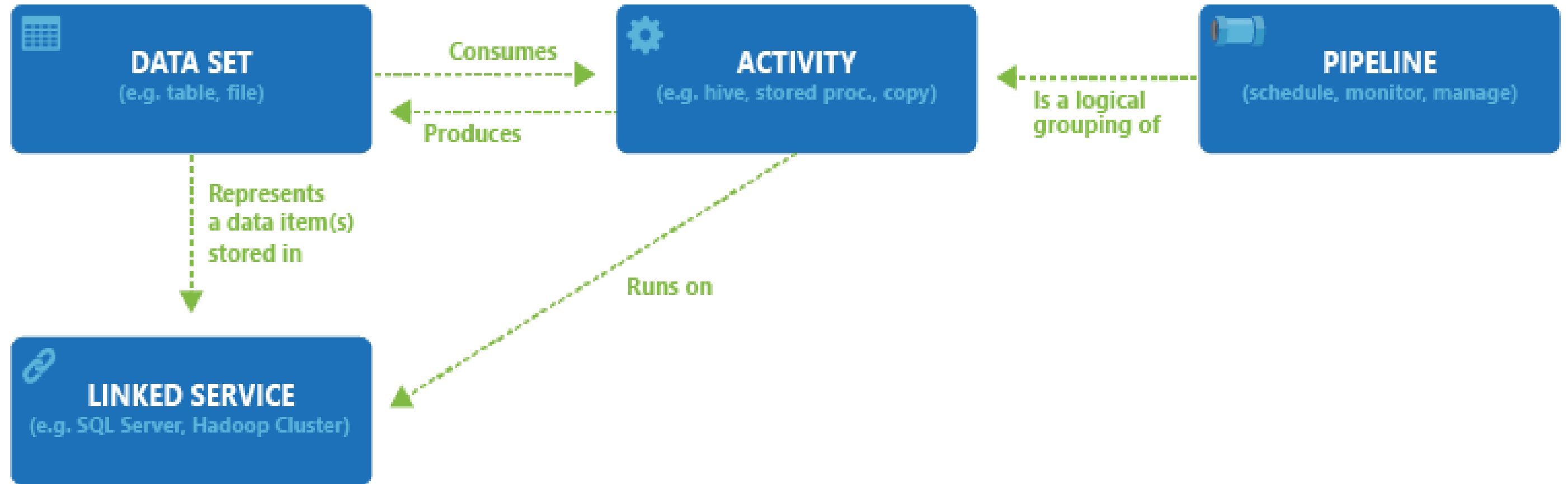
# Azure Data Factory

# 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



# Azure Data Factory Components



# 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](#)

## Give Feedback

- [UserVoice](#)

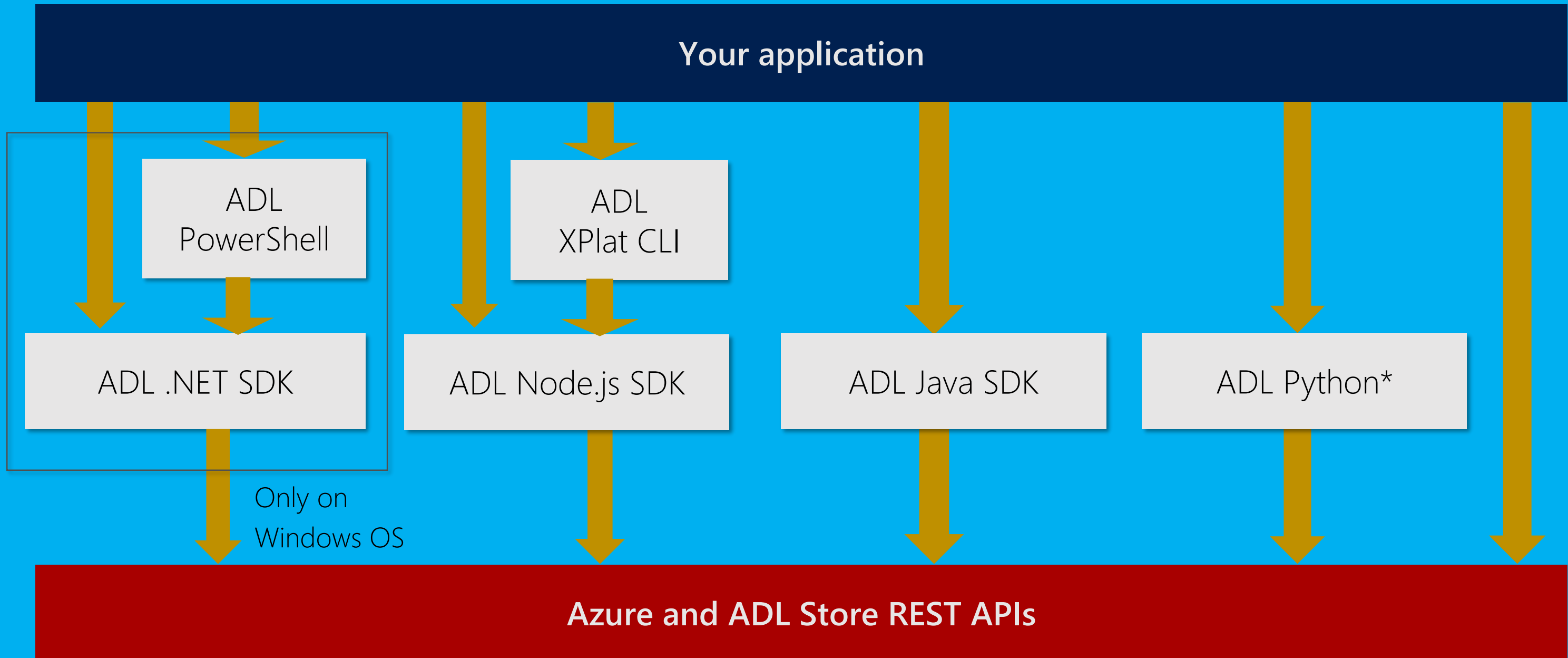
## Learn More

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

## Get Help

- [ADF Forums](#)

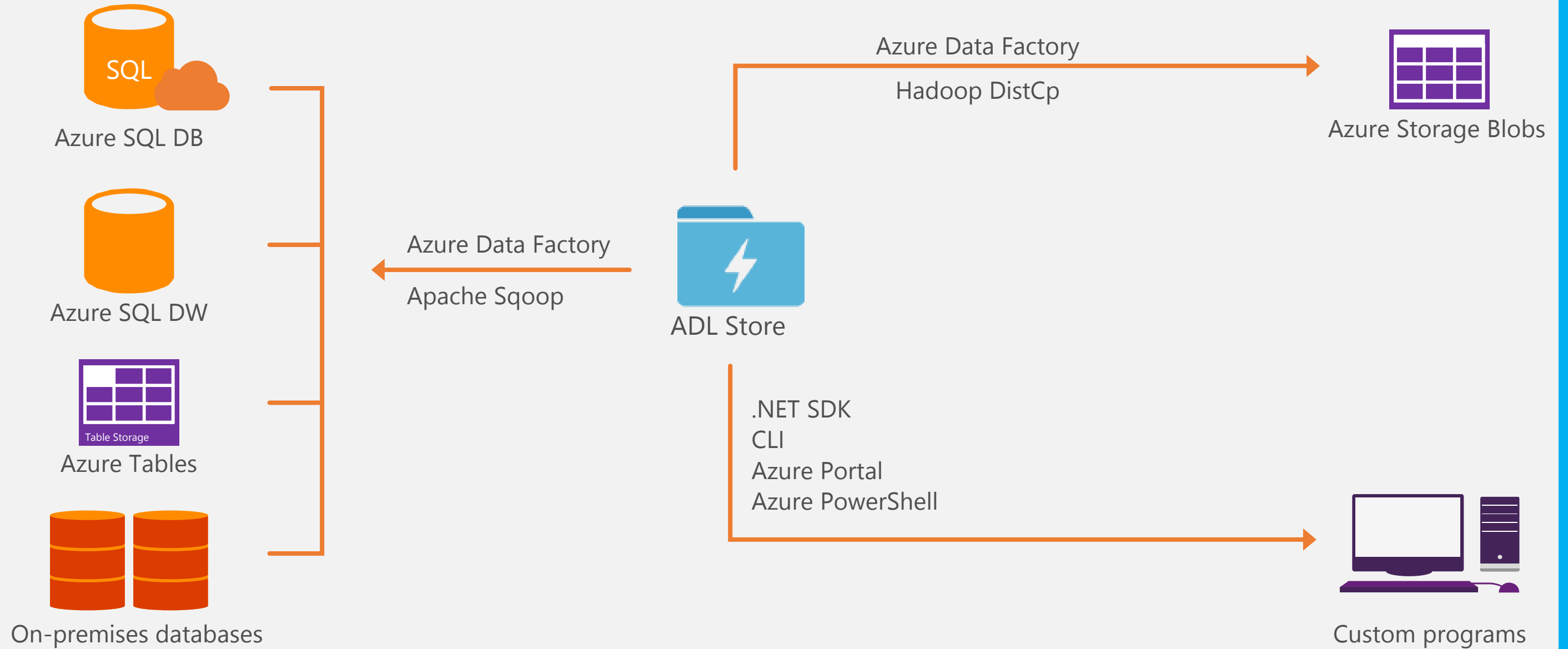
# Customizing using SDKs/APIs



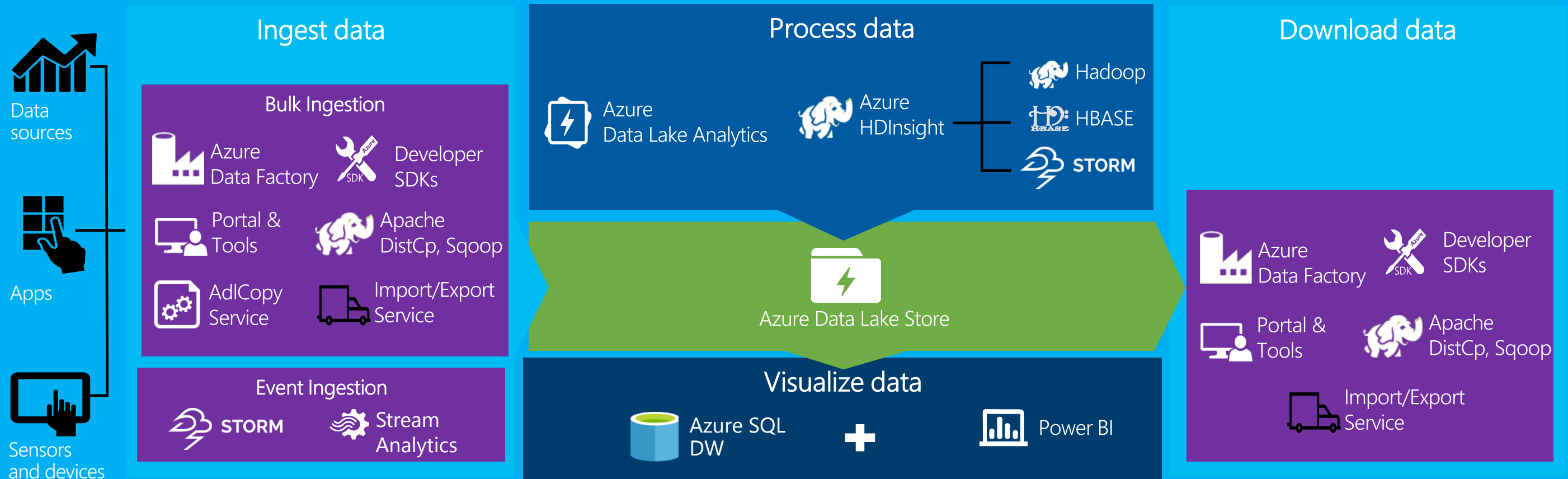
\* At General Availability

# Egress

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



# Big Data Analytics – Data Flow



# 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>



