# Batch Processing

Big data solutions often use long-running batch jobs to filter, aggregate, and otherwise prepare the data for analysis. Usually, these jobs involve reading source files from scalable storage (like HDFS, Azure Data Lake Store, and Azure Storage), processing them, and writing the output to new files in scalable storage.

The key requirement of such batch processing engines is the ability to scale out computations, in order to handle a large volume of data. Unlike real-time processing, however, batch processing is expected to have latencies (the time between data ingestion and computing a result) that measure in minutes to hours.

When to use this solution

Batch processing is used in a variety of scenarios, from simple data transformations to a more complete ETL (extract-transform-load) pipeline. In a big data context, batch processing may operate over very large data sets, where the computation takes significant time. (For example, see [Lambda architecture](https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/#lambda-architecture).) Batch processing typically leads to further interactive exploration, provides the modeling-ready data for machine learning, or writes the data to a data store that is optimized for analytics and visualization.

One example of batch processing is transforming a large set of flat, semi-structured CSV or JSON files into a schematized and structured format that is ready for further querying. Typically, the data is converted from the raw formats used for ingestion (such as CSV) into binary formats that are more performant for querying because they store data in a columnar format, and often provide indexes and inline statistics about the data.

## Architecture

* A batch processing architecture has the following logical components, shown in the diagram above.
* Data storage. Typically, a distributed file store that can serve as a repository for high volumes of large files in various formats. Generically, this kind of store is often referred to as a data lake.
* Batch processing. The high-volume nature of big data often means that solutions must process data files using long-running batch jobs to filter, aggregate, and otherwise prepare the data for analysis. Usually, these jobs involve reading source files, processing them, and writing the output to new files.
* Analytical data store. Many big data solutions are designed to prepare data for analysis and then serve the processed data in a structured format that can be queried using analytical tools.
* Analysis and reporting. The goal of most big data solutions is to provide insights into the data through analysis and reporting.
* Orchestration. With batch processing, typically some orchestration is required to migrate or copy the data into your data storage, batch processing, analytical data store, and reporting layers.

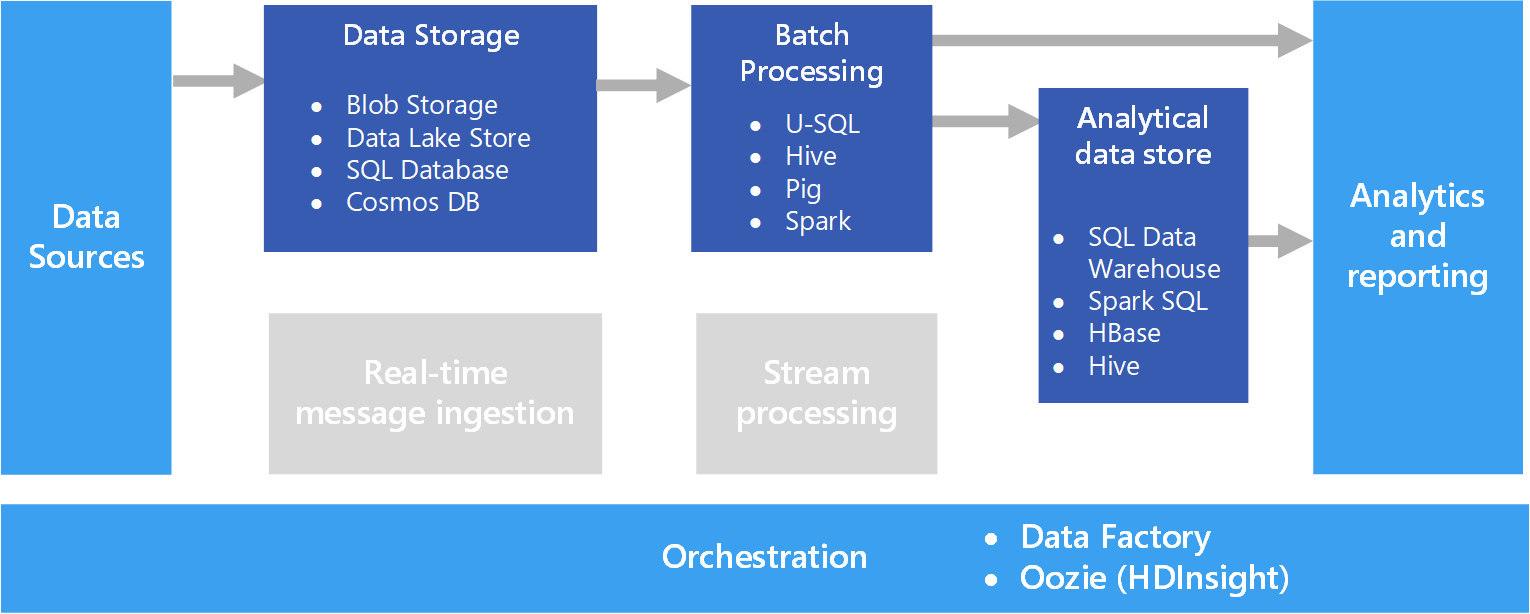
## Technology choices

1. Data Storage
2. Batch processing
3. Analytical and data store
4. Analytics and Reporting

## services that customers use to build data lakes for processing batch data

|  |  |  |  |
| --- | --- | --- | --- |
| **Ingestion** | **Processing/**  **Transformation** | **Analytics/Storage**            Azure files | **Visualization/**  **Machine Learning** |

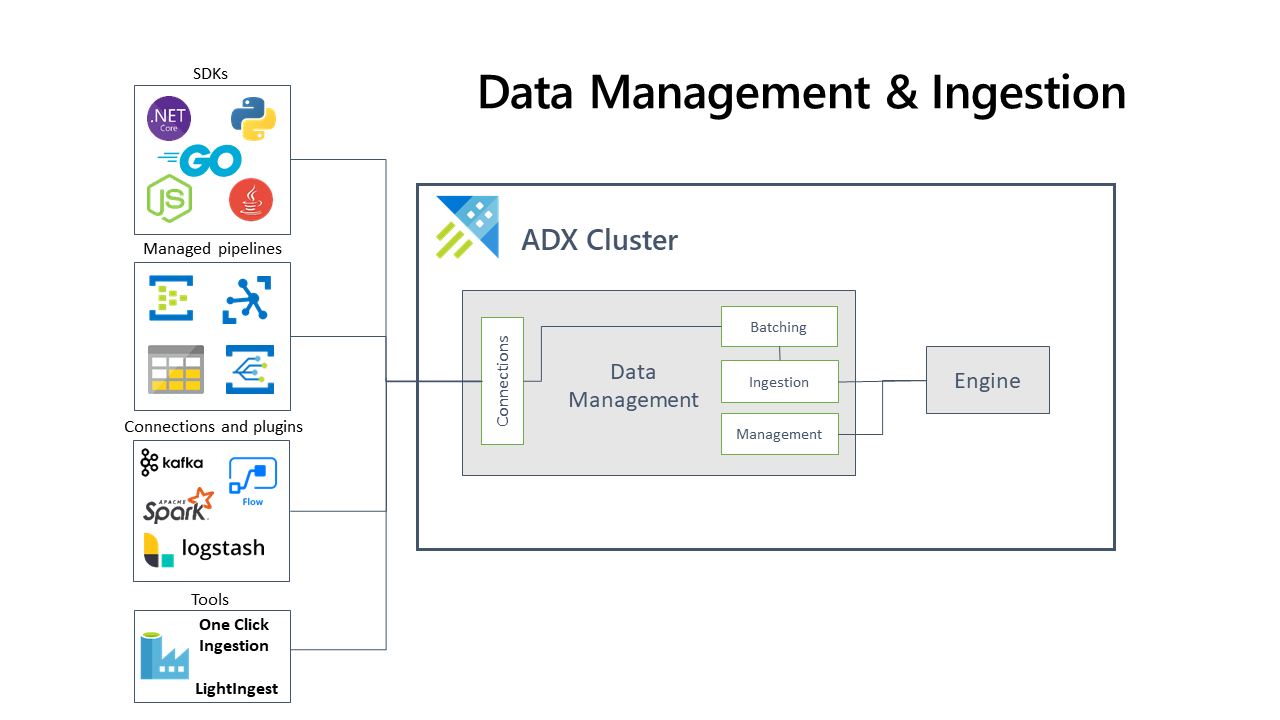
Below figure is a common example for batch processing in azure.



A common big data scenario is batch processing of data at rest. In this scenario, the source data is loaded into data storage, either by the source application itself or by an orchestration workflow. The data is then processed in-place by a parallelized job, which can also be initiated by the orchestration workflow. The processing may include multiple iterative steps before the transformed results are loaded into an analytical data store, which can be queried by analytics and reporting components.

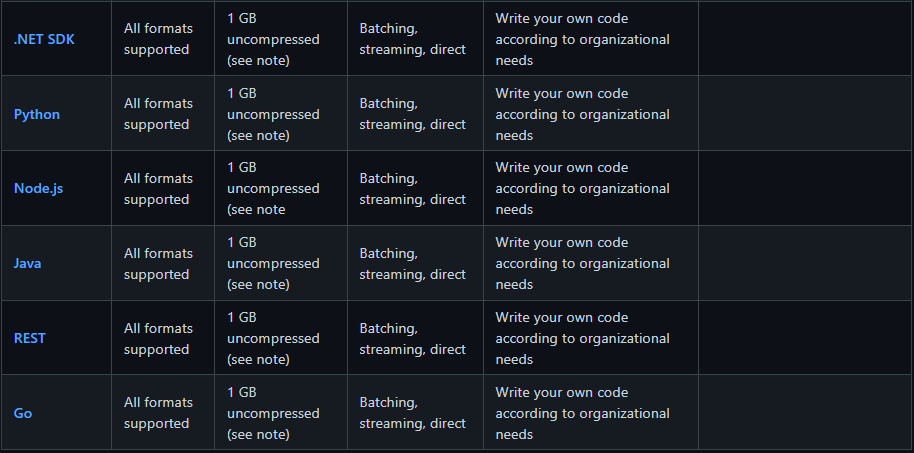
# Azure Data Explorer data ingestion overview

Data ingestion is the process used to load data records from one or more sources to import data into a table in Azure Data Explorer



## Comparing ingestion methods and tools





## [Controlling costs in Azure Data Explorer using down-sampling and aggregation](https://azure.microsoft.com/en-us/blog/controlling-costs-in-azure-data-explorer-using-down-sampling-and-aggregation/)

Azure Data Explorer (ADX) is an outstanding service for continuous ingestion and storage of high velocity telemetry data from cloud services and IoT devices. Leveraging its first-rate performance for querying billions of records, the telemetry data can be further analyzed for various insights such as monitoring service health, production processes, and usage trends. Depending on data velocity and retention policy, data size can rapidly scale to petabytes of data and increase the costs associated with data storage. A common solution for storage of large datasets for a long period of time is to store the data with differing resolution. The most recent data is stored at maximum resolution, meaning all events are stored in raw format. While the historic data is stored at reduced resolution, being filtered and/or aggregated. Above solution is often used for time series databases to control hot storage costs.

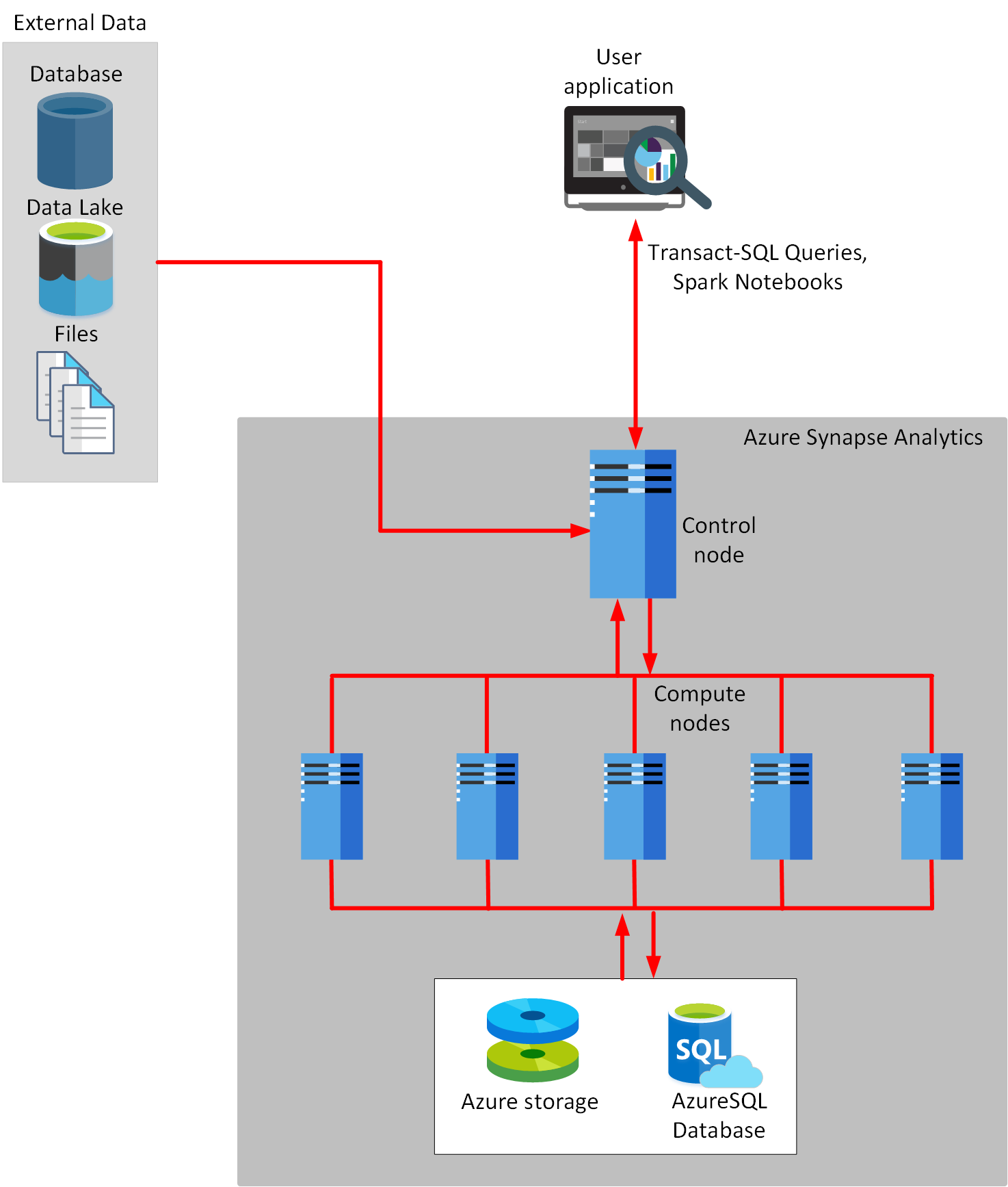
# Processing and transformation

Organizations generate data throughout their business. For analysis purposes, this data can be left in its raw, ingested format, or the data can be processed and saved to a specially designed data store or data warehouse. Azure enables businesses to implement either of these scenarios.

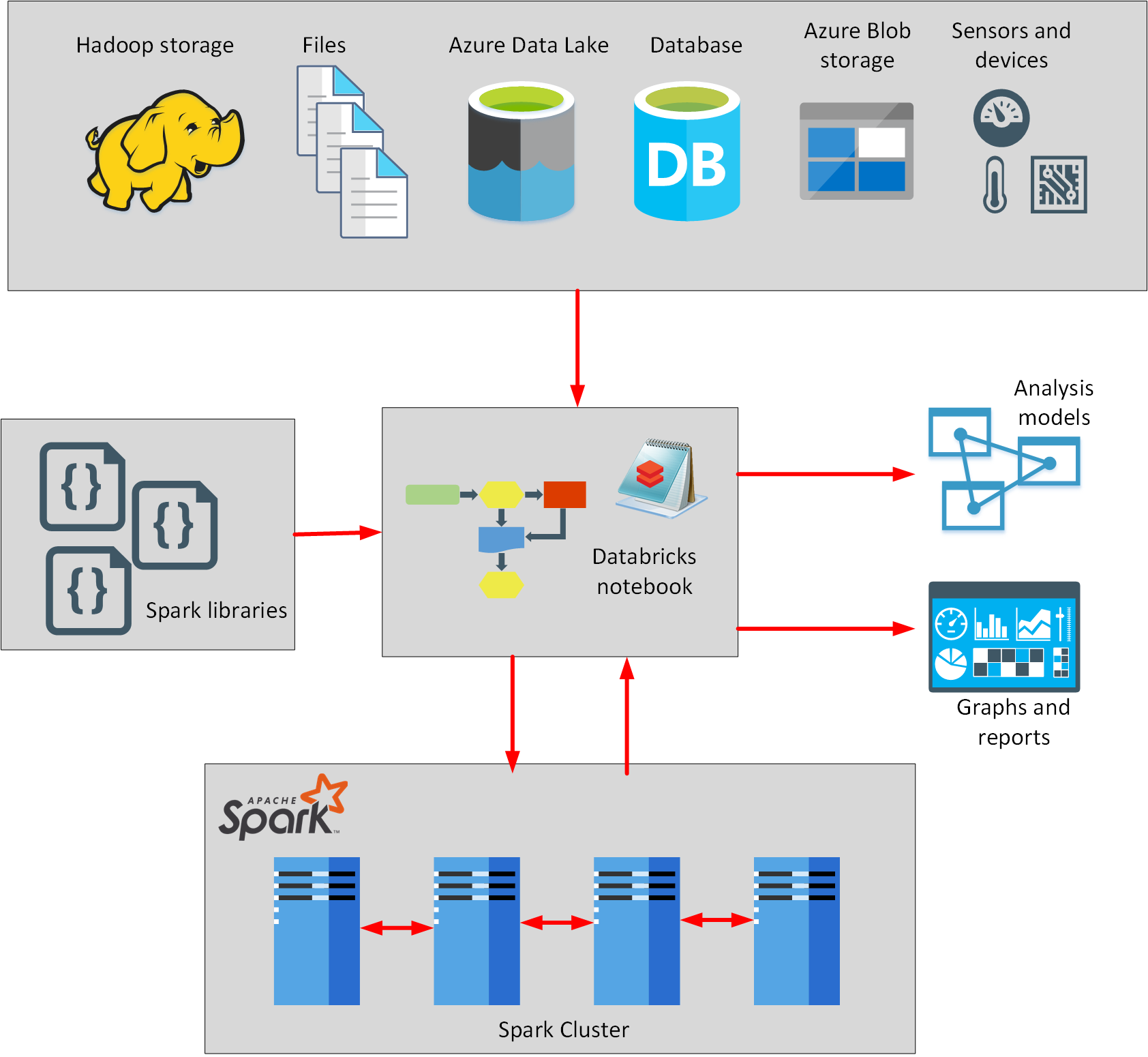
The most common options for data processing in azure

* Azure Databricks
* Azure Data Factory
* Azure Synapse Analytics
* Azure Data Lake

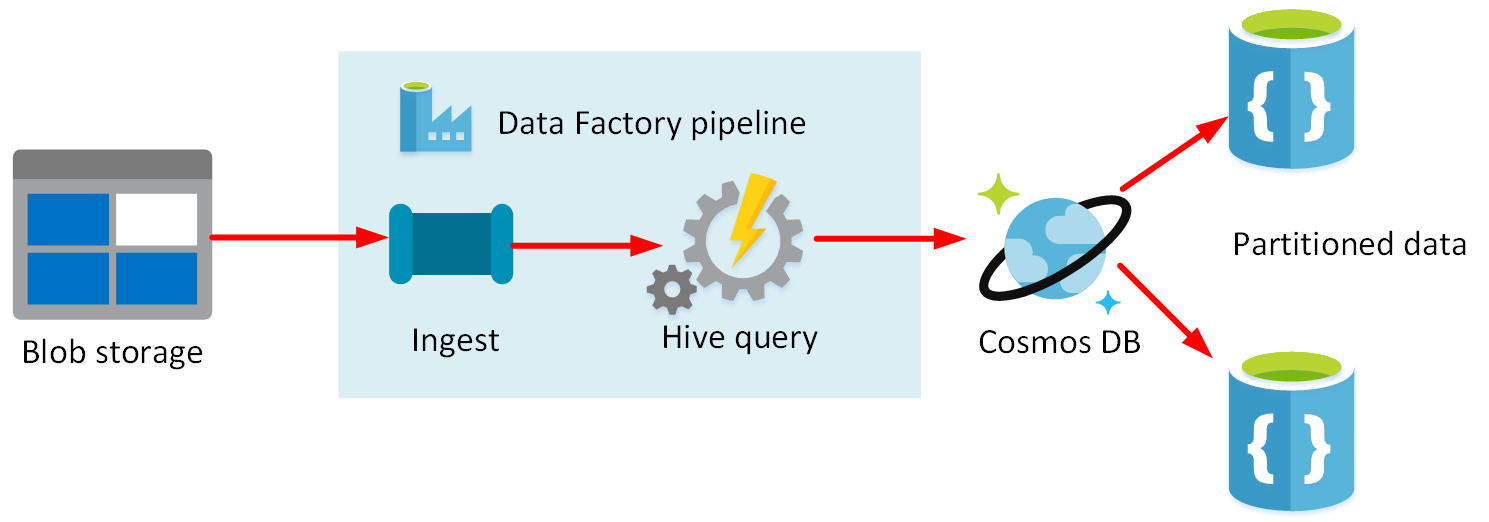
## Azure synapse



## Azure Databricks



## Azure Data Factory



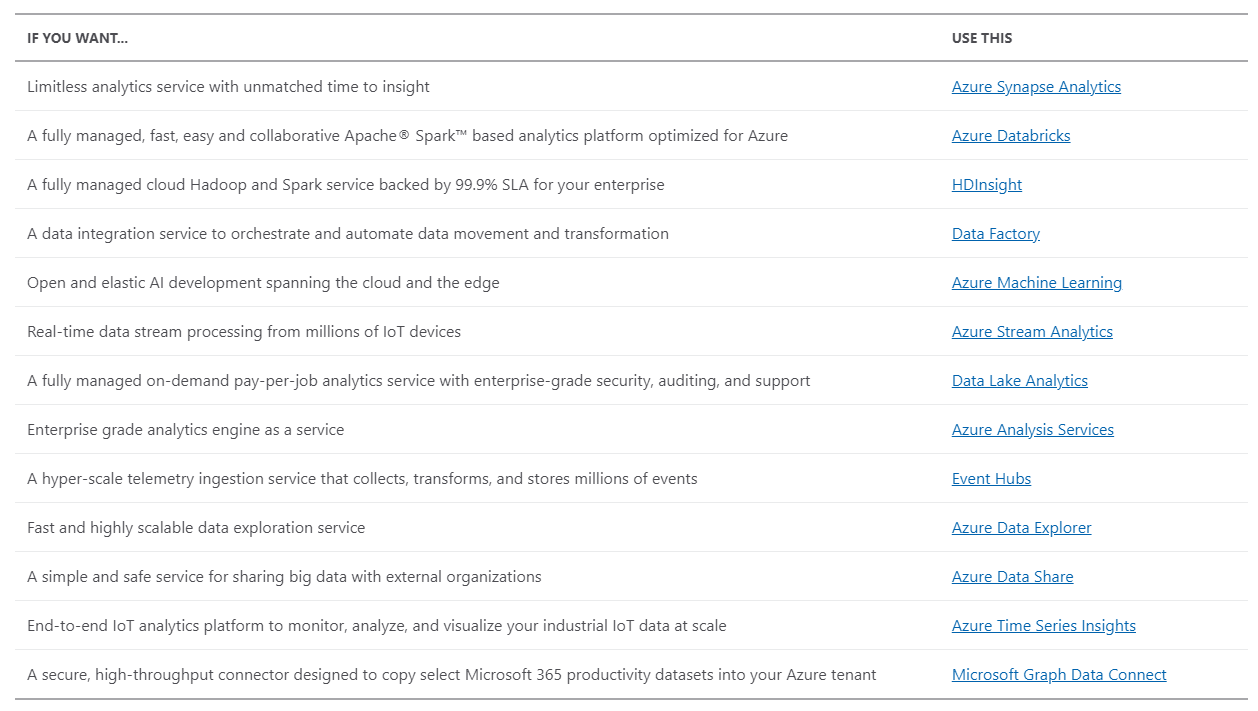
## Data Transformation

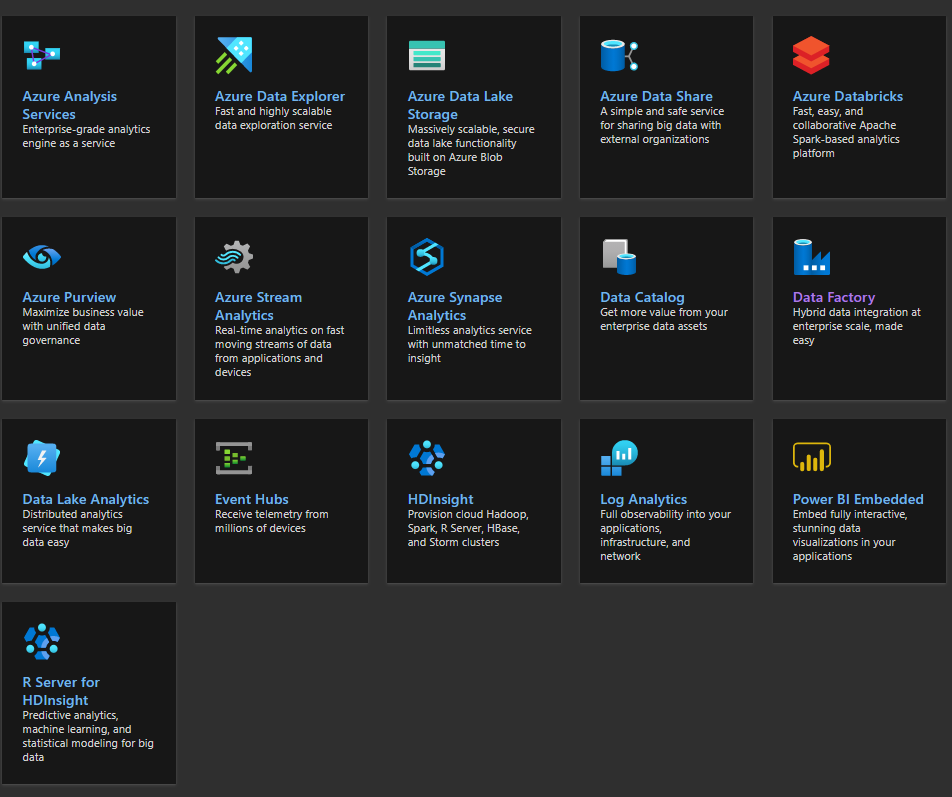
For the data transformation we can use below tools.

* Azure Data Factory
* Azure Synapse Analytics

# Analytics and core Azure Storage services

## Analytics Services





## Storage Services

Core storage services

The Azure Storage platform includes the following data services:

* [Azure Blobs](https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blobs-introduction): A massively scalable object store for text and binary data. Also includes support for big data analytics through Data Lake Storage Gen2.
* [Azure Files](https://docs.microsoft.com/en-us/azure/storage/files/storage-files-introduction): Managed file shares for cloud or on-premises deployments.
* [Azure Queues](https://docs.microsoft.com/en-us/azure/storage/queues/storage-queues-introduction): A messaging store for reliable messaging between application components.
* [Azure Tables](https://docs.microsoft.com/en-us/azure/storage/tables/table-storage-overview): A NoSQL store for schema less storage of structured data.
* [Azure Disks](https://docs.microsoft.com/en-us/azure/virtual-machines/managed-disks-overview): Block-level storage volumes for Azure VMs.

Each service is accessed through a storage account. To get started, see [Create a storage account](https://docs.microsoft.com/en-us/azure/storage/common/storage-account-create).

# Visualization and Machine Learning

## Visualization Services

Visualizations are very important thing get insights from data. Data visualization and reporting is a critical step in the data analytics process. Azure Data Explorer supports many BI services so you can use the one that best fits your scenario and budget.

## Power BI

Azure Data Explorer provides the capability to connect to [Power BI](https://powerbi.microsoft.com/) using various methods:

[Built-in native Power BI connector](https://docs.microsoft.com/en-us/azure/data-explorer/power-bi-connector)

[Query import from Azure Data Explorer into Power BI](https://docs.microsoft.com/en-us/azure/data-explorer/power-bi-imported-query)

[SQL query](https://docs.microsoft.com/en-us/azure/data-explorer/power-bi-sql-query)

## Microsoft Excel

Azure Data Explorer provides the capability to connect to [Microsoft Excel](https://products.office.com/excel) using the [built-in native Excel connector](https://docs.microsoft.com/en-us/azure/data-explorer/excel-connector), or [import a query](https://docs.microsoft.com/en-us/azure/data-explorer/excel-blank-query) from Azure Data Explorer into Excel.

## Grafana

[Grafana](https://grafana.com/) provides an Azure Data Explorer plugin that enables you to visualize data from Azure Data Explorer. You [set up Azure Data Explorer as a data source for Grafana, and then visualize the data](https://docs.microsoft.com/en-us/azure/data-explorer/grafana).

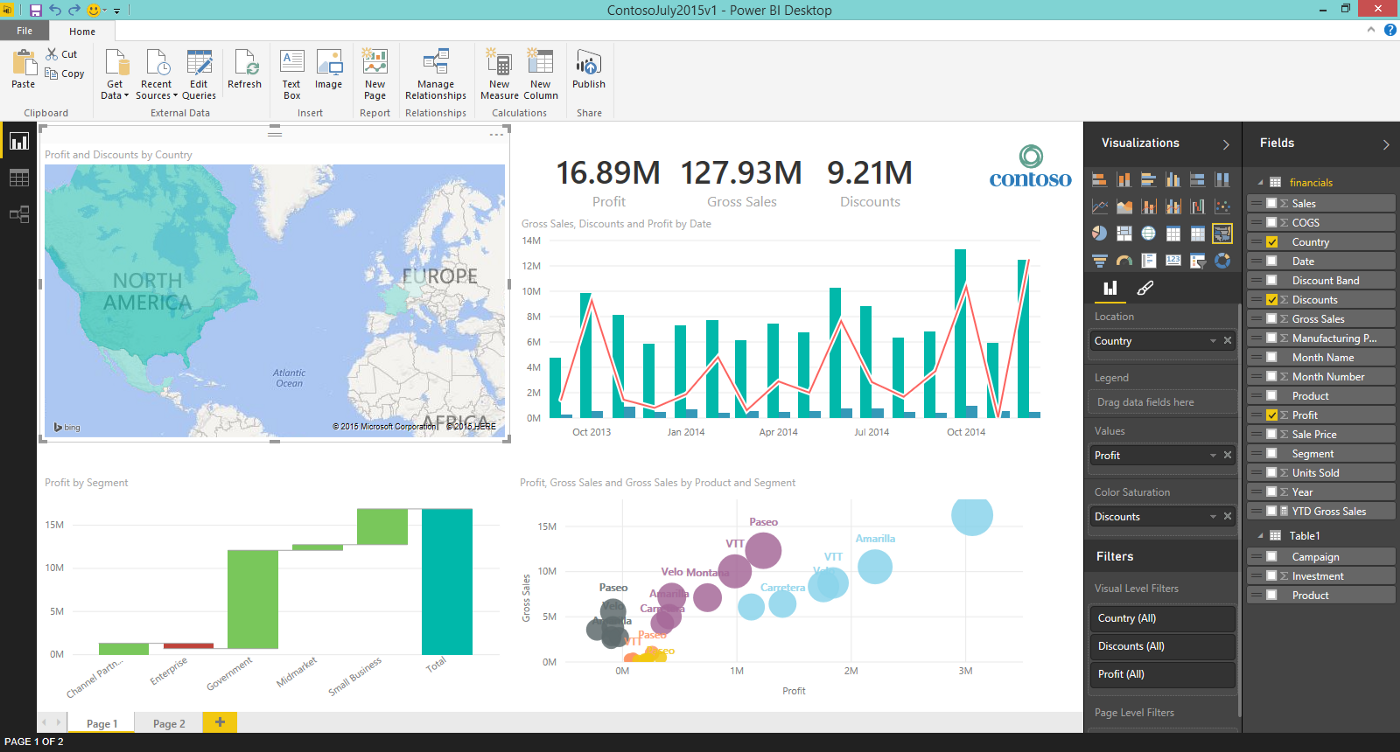
## Kibana

Azure Data Explorer provides the capability to connect to [Kibana](https://www.elastic.co/guide/en/kibana/6.8/discover.html) using K2Bridge, an open source connector. You [set up Azure Data Explorer as a data source for Kibana, and then visualize the data](https://docs.microsoft.com/en-us/azure/data-explorer/k2bridge).

## Tableau

Azure Data Explorer provides the capability to connect to [Tableau](https://www.tableau.com/) using the [ODBC connector](https://docs.microsoft.com/en-us/azure/data-explorer/connect-odbc) and then [visualize the data in Tableau](https://docs.microsoft.com/en-us/azure/data-explorer/tableau).

Below figure is an example for Visualizing Data with Azure Databricks and Power BI Desktop.

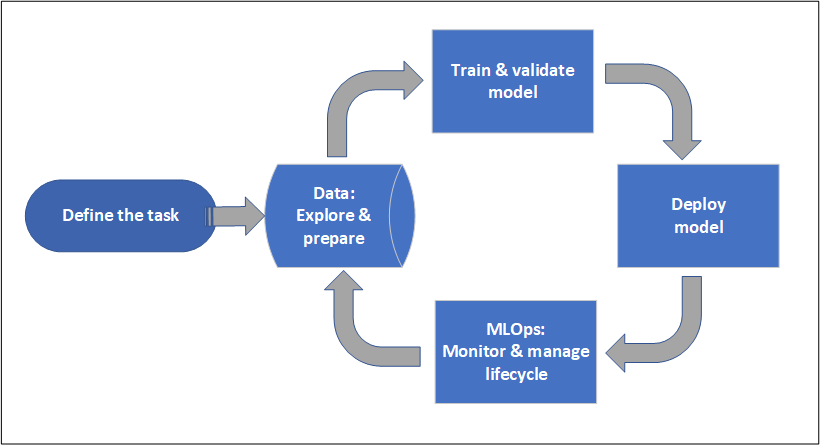


## Machine Learning

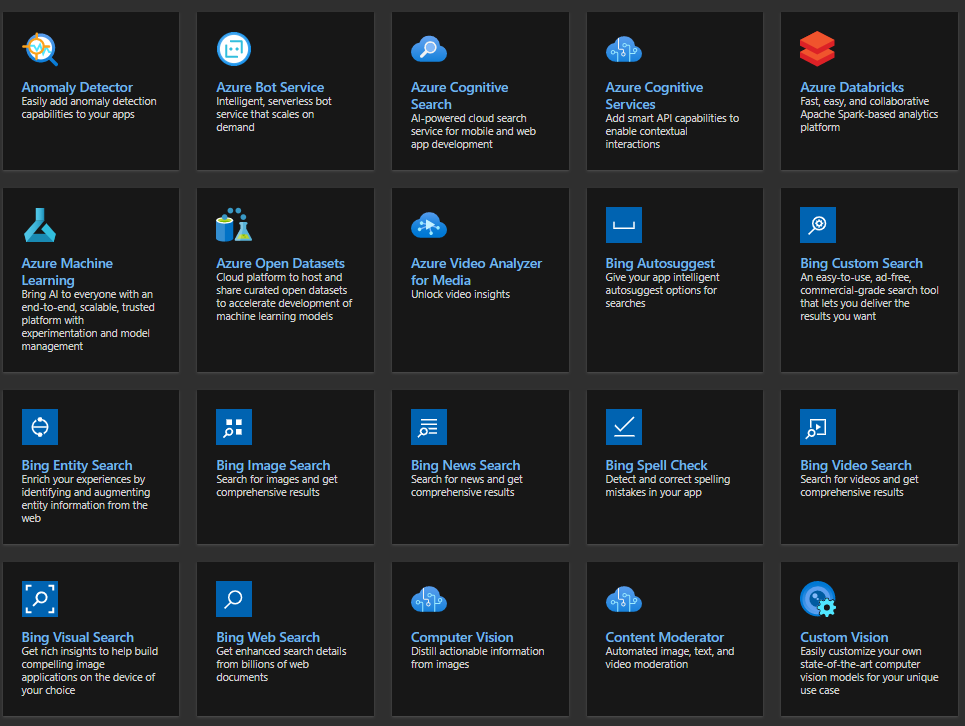
Azure Machine Learning is a cloud service for accelerating and managing the machine learning project lifecycle. Machine learning professionals, data scientists, and engineers can use it in their day-to-day workflows: Train and deploy models and manage MLOps.

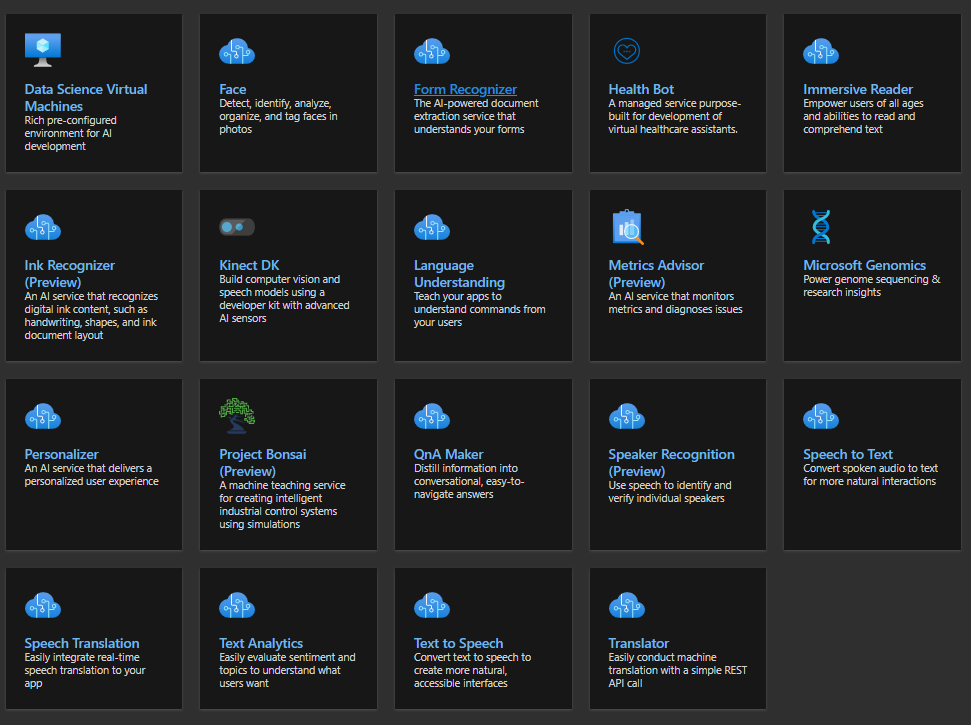
You can create a model in Azure Machine Learning or use a model built from an open-source platform, such as Pytorch, TensorFlow, or scikit-learn. MLOps tools help you monitor, retrain, and redeploy models.

## Machine learning project workflow



## Azure Ai-Machine Learning services





## References

Docs.microsoft.com. 2021. *Azure documentation*. [online] Available at: <https://docs.microsoft.com/en-us/azure/?product=all> [Accessed 10 August 2021].