Environment Setup

1. Azure Subscription
2. Data Factory
   1. Ingestion and Transfer of data
   2. Data flow is part of DF
3. Blob Storage account
   1. To keep Data
   2. To keep config files
   3. To keep Scripts for HD insight Transformation
4. Data Lake Storage Gen2
   1. To keep ingested and Transformed Data
5. Azure Sql data base
   1. Storage for Reporting
6. Azure Databricks Cluster
   1. For doing transformation
7. Azure HD insight Cluster
   1. For doing transformation

Note – Databricks cluster/HD insight Cluster are expensive so we will create when we need then destroy asap

What comes with Free account

1. VM- 750 hours
2. Blob Storage – 5GB
3. SQL DB – 250GB

What is ADF

Azure Data Factory is a cloud-based ETL and data integration service that enables you to create data-driven workflows to:

* Orchestrate data movement.
* Transform data at scale.

Data-driven workflows are also known as pipelines.

A package is similar to an Azure Data Factory pipeline. Each package defines a process to extract, load, transform, or otherwise work with data.

Azure provides several technologies that you can implement to help with your organization's data analytics. These include:

* Azure Synapse Analytics
* Azure Blob Storage
* Azure Data Lake Storage
* Azure Data Lake Analytics
* Azure Analysis Services
* Azure HDInsight
* Azure Databricks
* Azure Machine Learning

You can use some or all of these services, as needed, to analyze your organization's data. However, none of these services addresses data integration. *Data integration* enables you to collect data from multiple sources, and then load this combined data to a location suitable for data analysis. If necessary, you can transform the data during this process. Although you can perform these tasks manually, you could consider using Azure Data Factory.

Big data community

For them, Azure Data Factory provides a means to create and run pipelines in the cloud. These pipelines can access both cloud and on-premises data services. These pipelines typically work with technologies such as Azure Synapse Analytics, Azure Blobs, Azure Data Lake, Azure HDInsight, Azure Databricks, and Azure Machine Learning

Relational data warehousing community

This community typically relies on technologies such as Microsoft SQL Server. SQL Server Integration Services (SSIS) is often used to create SSIS packages. For this community, Azure Data Factory provides the ability to run SSIS packages on Azure, thereby letting them access both cloud and on-premises data services.

**How Azure Data Factory can help with data analytics**

By using Azure Data Factory, you can:

* Build complex ETL processes. These processes can transform data visually by using either data flows or compute services such as:
  + Azure HDInsight Hadoop
  + Azure Databricks
  + Azure SQL Database
* Publish this transformed data to data stores for business intelligence apps to consume.

Application, timeline

Description automatically generated with medium confidence

# How Azure Data Factory works

Azure Data Factory functions:

* Connect and collect
* Transform and enrich
* Continuous integration and delivery (CI/CD) and publish
* Monitoring

You'll also learn about the key components of Azure Data Factory. These are:

* Pipelines
* Activities
* Datasets
* Linked services
* Data flows
* Integration runtimes

## Azure Data Factory functions

Azure Data Factory consists of several functions that combine to provide your data engineers with a complete data analytics platform.

### Connect and collect

1. collect the required data from on-premises sources and in the cloud. The data might be:

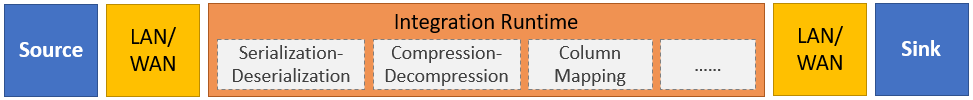
* Structured
* Unstructured
* Semi-structured

In addition, this disparate data might arrive at different speeds and intervals.

The copy activity performs the following high-level steps:

1. Read data from source data store.
2. Perform the following tasks on the data:
   * Serialization/deserialization
   * Compression/decompression
   * Column mapping
3. Write data to the destination data store (known as the sink).

This process is summarized in the following graphic.



### Transform and enrich

After you've successfully copied the data to a central cloud-based location, you can process and transform the data as needed.

You'll use Azure Data Factory mapping data flows to achieve this.

Data flows enable you to create data transformation graphs that run on Spark. However, you don't need to understand Spark clusters or Spark programming.

**Tip**

Although not necessary, you might prefer to code your transformations manually. If so, Azure Data Factory supports external activities for running your transformations.

### CI/CD and publish

Support for CI/CD enables you to develop and deliver your ETL processes incrementally before you publish. Azure Data Factory provides for CI/CD of your data pipelines by using:

* Azure DevOps
* GitHub

**Note**

Continuous integration means automatically testing each change made to your codebase as soon as possible.

Continuous delivery follows this testing and pushes changes to a staging or production system.

After Azure Data Factory has refined the raw data, you can load the data into whichever analytics engine your business users can access from their business intelligence tools, including:

* Azure Synapse Analytics
* Azure SQL Database
* Azure Cosmos DB

### Monitor

After you've successfully built and deployed your data integration pipeline, it's important that you can monitor your scheduled activities and pipelines. This enables you to track success and failure rates. Azure Data Factory provides support for pipeline monitoring by using one of the following:

* Azure Monitor
* API
* PowerShell
* Azure Monitor logs
* Health panels in the Azure portal

**Azure Data Factory components**

Azure Data Factory is comprised of the components described in the following table.

| **AZURE DATA FACTORY COMPONENTS** | |
| --- | --- |
| **Component** | **Description** |
| Pipelines | A logical grouping of activities that perform a specific unit of work. These activities together perform a task. The advantage of using a pipeline is that you can more easily manage the activities as a set instead of as individual items. |
| Activities | A single processing step in a pipeline. Azure Data Factory supports three types of activity: data movement, data transformation, and control activities. |
| Datasets | Represent data structures within your data stores. These point to (or reference) the data that you want to use in your activities as either inputs or outputs. |
| Linked services | Define the required connection information needed for Azure Data Factory to connect to external resources, such as a data source. Azure Data Factory uses these for two purposes: to represent a **data store** or a **compute resource**. |
| Data flows | Enable your data engineers to develop data transformation logic without needing to write code. Data flows are run as activities within Azure Data Factory pipelines that use scaled-out Apache Spark clusters. |
| Integration runtimes | Azure Data Factory uses the compute infrastructure to provide the following data integration capabilities across different network environments: data flow, data movement, activity dispatch, and SSIS package execution. In Azure Data Factory, an integration runtime provides the bridge between the activity and linked services. |

As indicated in the following graphic, these components work together to provide a complete end-to-end platform for data engineers. By using Data Factory, you can:

* Set triggers on-demand and schedule data processing based on your needs.
* Associate a pipeline with a trigger, or manually start it as and when needed.
* Connect to linked services, such as on-premises apps and data, or Azure services via integration runtimes.
* Monitor all of your pipeline runs natively in the Azure Data Factory user experience, or by using Azure Monitor.

