Azure Data Factory

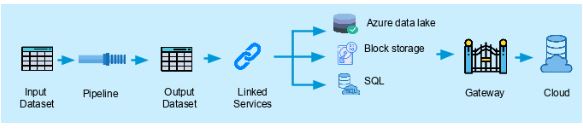
Organizations often face a situation where their data generation from applications or products increases exponentially. As the data is generated from different products, it is difficult to analyze and store all of the [data](https://www.simplilearn.com/what-is-data-article).

Azure Data Factory can help to manage such data. It stores all kinds of data with the help of data lake storage. You can then analyze the data and transform it using pipelines, and finally publish the organized data and visualize it with third-party applications, like Apache Spark or [Hadoop](https://www.simplilearn.com/tutorials/hadoop-tutorial/what-is-hadoop).

## **What is Azure Data Factory?**

Azure Data Factory is a cloud-based integration service that orchestrates and automates the movement and transformation of data. It works heavily on the data that you store.

Let us discuss the process followed in the Azure Data Factory.



### **Input Datasets**

This represents the collection of data within the data stores. The data passes through a pipeline for processing.

#### **Stand Out From Your Peers this Appraisal Season**

Start Learning With Our FREE Courses[ENROLL NOW](https://www.simplilearn.com/skillup-free-online-courses?utm_source=frs&utm_medium=skillup-home-banner&utm_campaign=frs-tutorial-skillup-home-promotion)



### **Pipeline**

A pipeline consists of a group of activities, such as:

* Data movement activity
* Data transformation activity using:
  + [SQL](https://www.simplilearn.com/sql-database-training-course)
  + Stored procedures
  + Hive

### **Output Datasets**

After the data is transformed into the pipeline, we get an output dataset. Here, we get a structured form of data.

### **Linked Services**

The data from output datasets passes to linked services, such as:

* Azure Data Lake
* Block storage
* SQL

Linked services contain information needed to connect to external sources. This is similar to the concept of a connection string in an SQL Server, where you define the source and destination of your data.

### **Gateway**

This connects your on-premises data to the cloud. It consists of a client agent that is installed on the on-premises data system, which then connects to the Azure data.

### **Cloud**

The data is analyzed and visualized using a number of analytical frameworks, like [Apache Spark](https://www.simplilearn.com/big-data-and-analytics/apache-spark-scala-certification-training), R, Hadoop, and so on.

## **What is Azure Data Lake?**

[Azure Data Lake](https://azure.microsoft.com/en-us/solutions/data-lake/) is a highly scalable, distributed, parallel file system in the cloud that is specifically designed to work with multiple analytics frameworks.

The data in output datasets (collected from mobile, the web, social platforms, etc.) is sent into the Azure Data Lake Store. It is then provided to external frameworks, like R and Apache Spark.

Data Lake works on two main concepts: storage and analytics.

### **Storage**

Storage is unlimited, allowing users to save very large files. A variety of data (like unstructured or structured data) can be stored here.

### **Analytics**

Through analytics, you can monitor and diagnose real-time data from connected devices, such as vehicles, buildings, or machinery to initiate actions such as generating alerts, responding to events, and optimizing operations.

You can also monitor financials such as:

* Financial transactions in real-time to detect fraudulent activity
* The use of a credit card across geographic locations
* The number of transactions on a single credit card