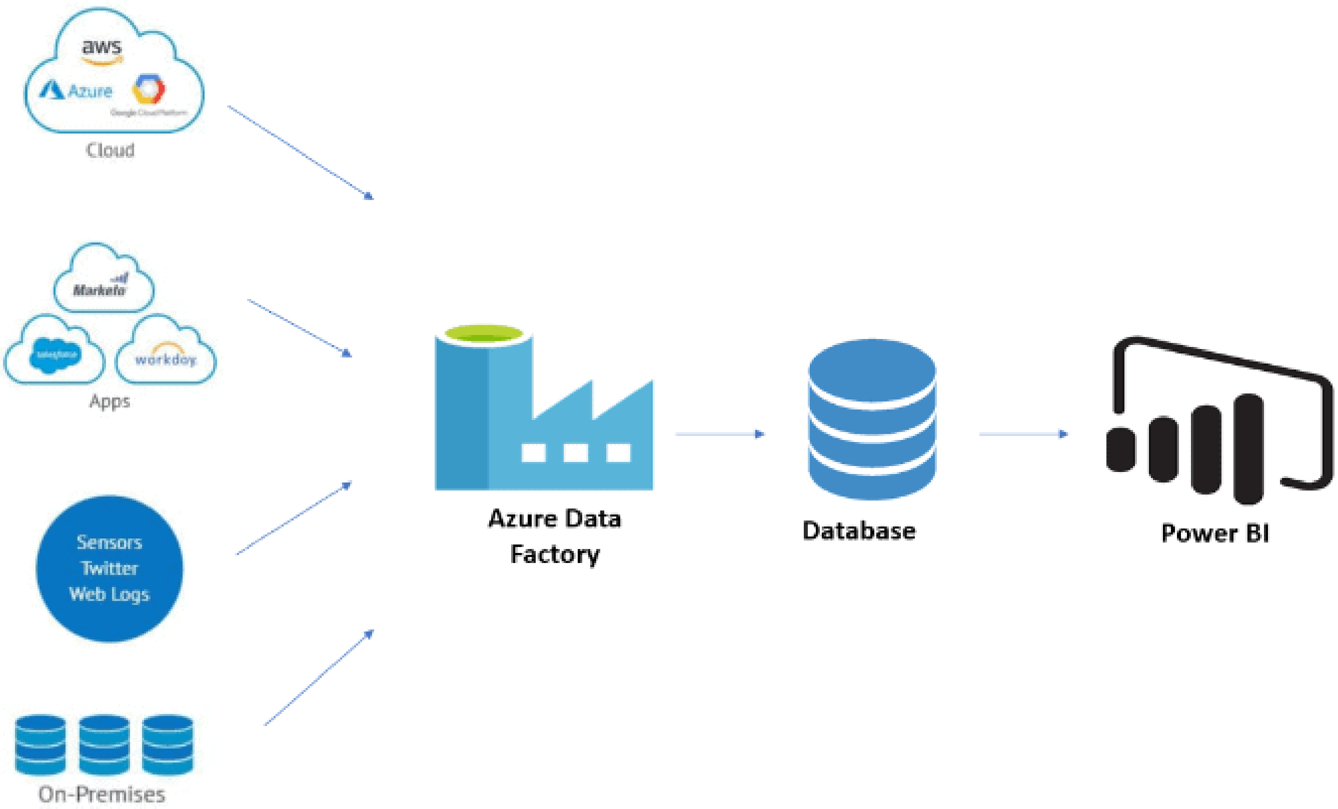
Azure Data Factory for Beginners

Azure Data Factory is a **cloud-based ETL** and **data integration service** that allows us to create data-driven pipelines for orchestrating data movement and transforming data at scale.

In this blog, we'll learn about the **Microsoft Azure Data Factory (ADF)** service. This service permits us to combine data from multiple sources, reformat it into analytical models, and save these models for following querying, visualization, and reporting.

What Is ADF?

- ADF is defined as a **data integration service**.
- The aim of ADF is to fetch data from one or more data sources and convert them into a format that we process.
- The data sources might contain noise that we need to filter out. ADF connectors enable us to pull the interesting data and remove the rest.
- ADF to ingest data and load the data from a variety of sources into Azure Data Lake Storage.
- It is the **cloud-based ETL** service that allows us to create data-driven pipelines for **orchestrating** data movement and transforming data at scale.



What Is a Data Integration Service?

- Data integration involves the collection of data from one or more sources.
- Then includes a process where the data may be transformed and cleansed or may be augmented with additional data and prepared.

- Finally, the combined data is stored in a data platform service that deals with the type of analytics that we want to perform.
- This process can be automated by ADF in an arrangement known as **Extract, Transform, and Load (ETL)**.

What Is ETL?

1) Extract

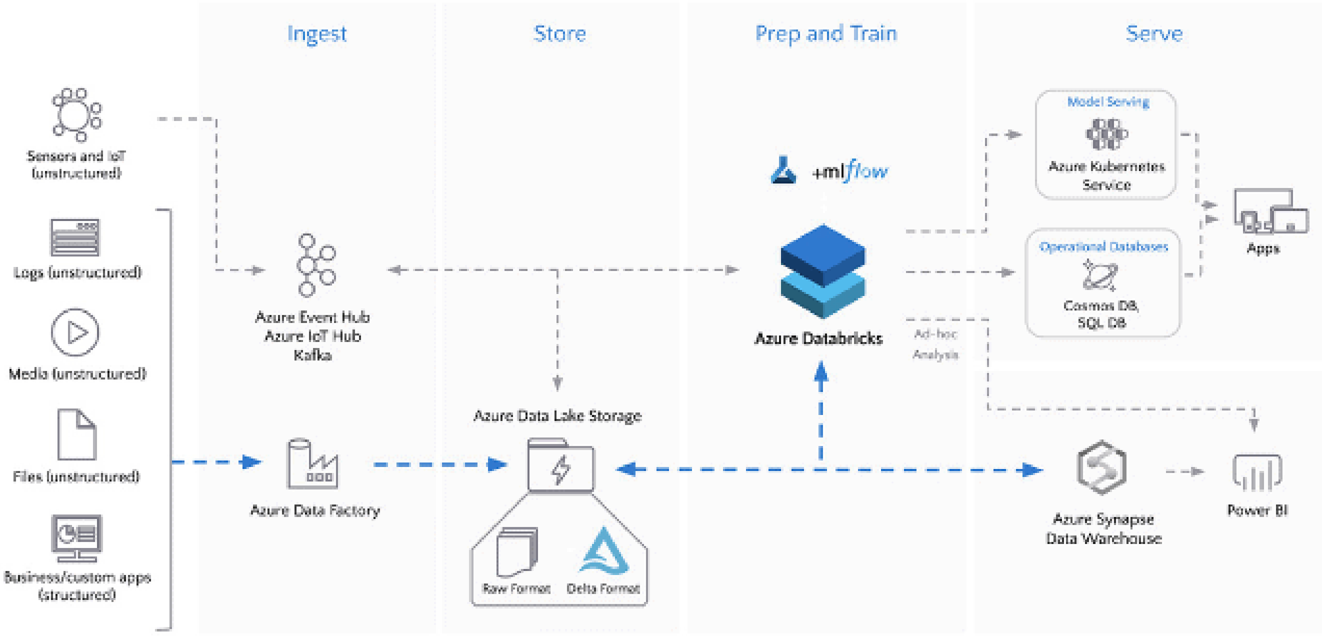
- In this extraction process, data engineers define the data and its source.
- **Data source**: Identify source details such as the subscription, resource group, and identity information such as secretor a key.
- **Data**: Define data by using a set of files, a database query, or an Azure Blob storage name for blob storage.

2) Transform

- Data transformation operations can include combining, splitting, adding, deriving, removing, or pivoting columns.
- Map fields between the data destination and the data source.

3) Load

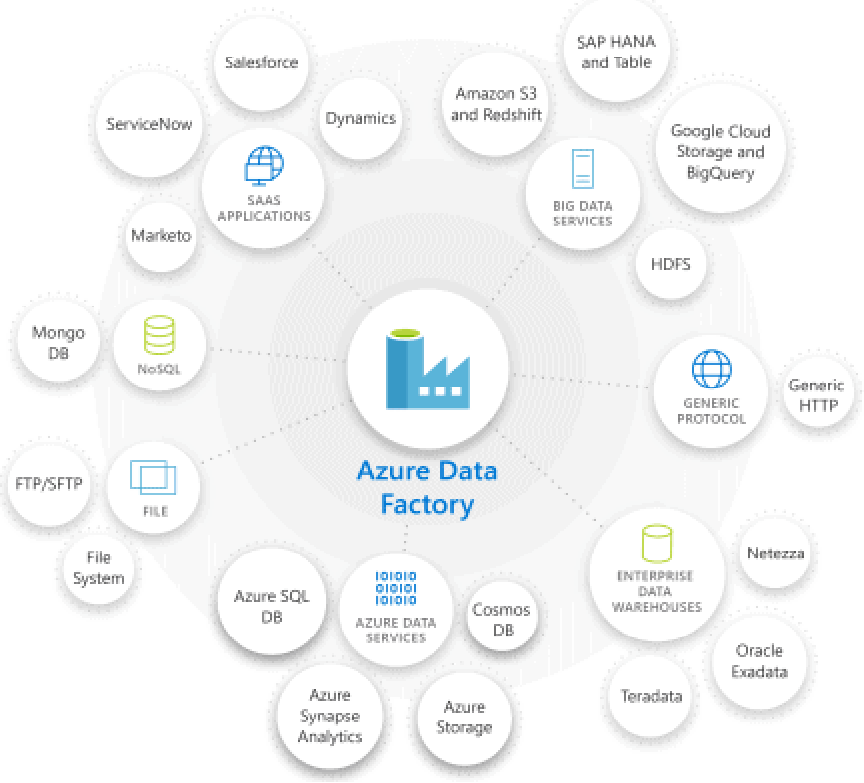
- During a load, many Azure destinations can take data formatted as a file, JavaScript Object Notation (JSON), or blob.
- Test the ETL job in a test environment. Then shift the job to a production environment to load the production system.



Go through this Microsoft Azure Blog to get a clear understanding of **Azure SQL**

4) ETL tools

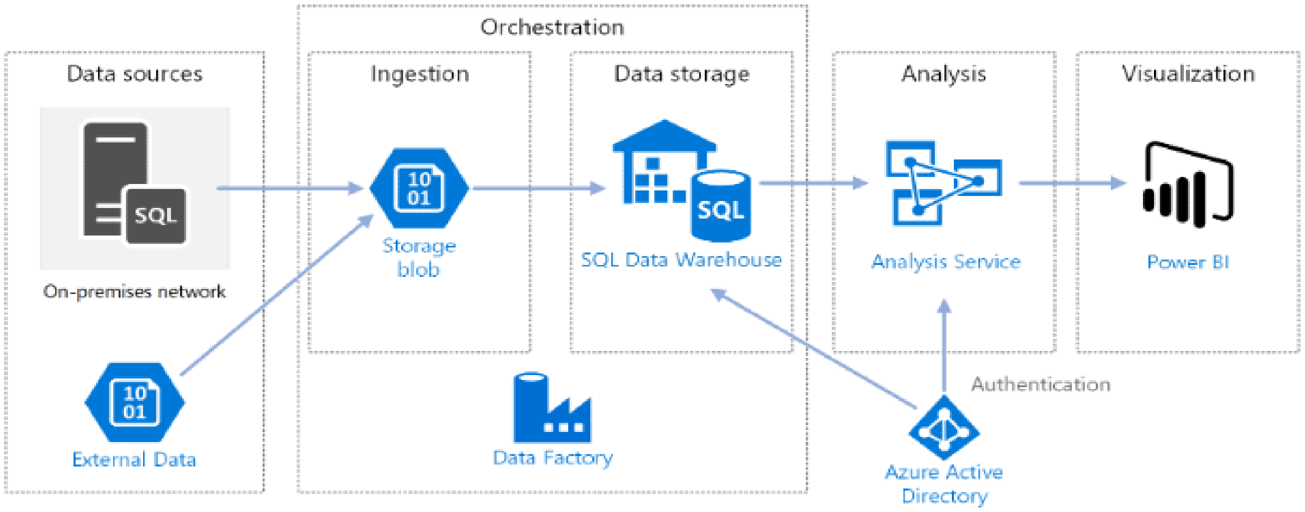
- Azure Data Factory provides approximately 100 enterprise connectors and robust resources for both code-based and code-free users to accomplish their data transformation and movement needs.



Also read: How Azure Event Hub & Event Grid Works?

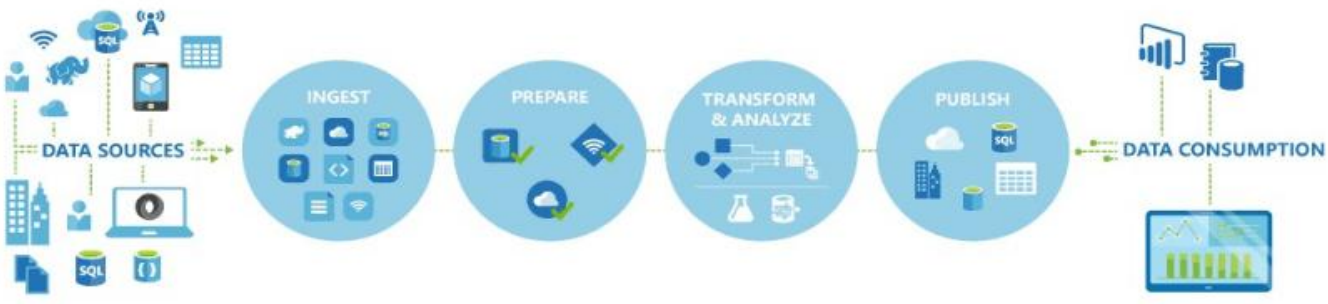
What Is Meant By Orchestration?

- Sometimes ADF will instruct another service to execute the actual work required on its behalf, such as a Databricks to perform a transformation query.
- ADF hardly orchestrates the execution of the query and then prepare the pipelines to move the data onto the destination or next step.



Copy Activity In ADF

- In ADF, we can use the Copy activity to copy data between data stores located on-premises and in the cloud.
- After we copy the data, we can use other activities to further transform and analyze it.
- We can also use the DF Copy activity to publish transformation and study results for business intelligence (BI) and application consumption.



1) Monitor Copy Activity

- Once we've created and published a pipeline in ADF, we can associate it with a trigger.
- We can monitor all of our pipelines runs natively in the ADF user experience.
- To monitor the Copy activity run, go to your DF **Author & Monitor** UI.
- On the **Monitor** tab page, we see a list of the pipeline runs, click the **pipeline name** link to access the list of activity runs in the pipeline run.

2) Delete Activity In ADF

- Back up your files before you are deleting them with the **Delete activity** in case you wish to restore them in the future.
- Make sure that Data Factory has to write permissions to delete files or folders or from the storage store.

To Know More About **Azure Databricks** click here

How ADF work?

1) Connect and Collect

- Enterprises have data of various types such as structured, unstructured, and semi-structured.
- The first step collects all the data from a different source and then move the data to a centralized location for subsequent processing.
- We can use the Copy Activity in a data pipeline to move data from both cloud source and on-premises data stores to a centralized data store in the cloud.

2) Transform and Enrich

- After data is available in a centralized data store in the cloud, transform, or process the collected data by using ADF mapping data flows.
- ADF supports external activities for executing our transformations on compute services such as **Spark, HDInsight Hadoop, Machine Learning, Data Lake Analytics.**

3) CI/CD and Publish

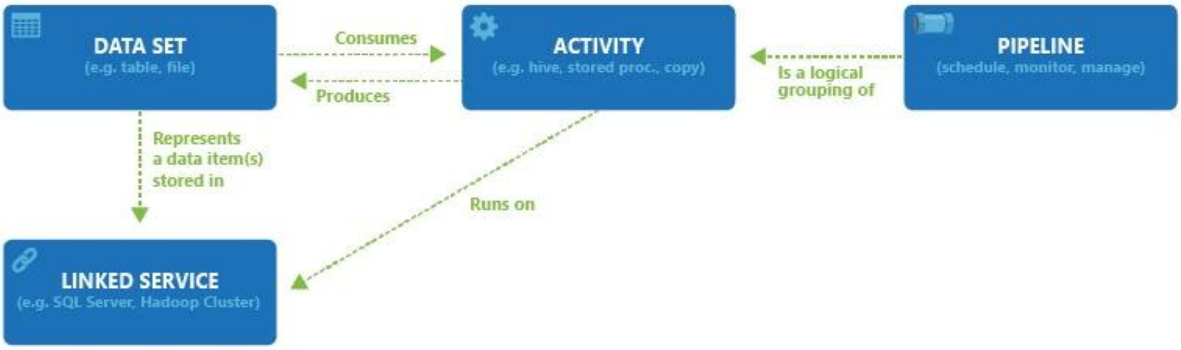
- ADF offers full support for CI/CD of our data pipelines using GitHub and Azure DevOps.
- After the raw data has been refined, ad the data into **Azure SQL Database, Azure Data Warehouse, Azure CosmosDB**

4) Monitor

- ADF has built-in support for pipeline monitoring via Azure Monitor, PowerShell, API, Azure Monitor logs, and health panels on the Azure portal.

5) Pipeline

- A pipeline is a logical grouping of activities that execute a unit of work. Together, the activities in a pipeline execute a task.

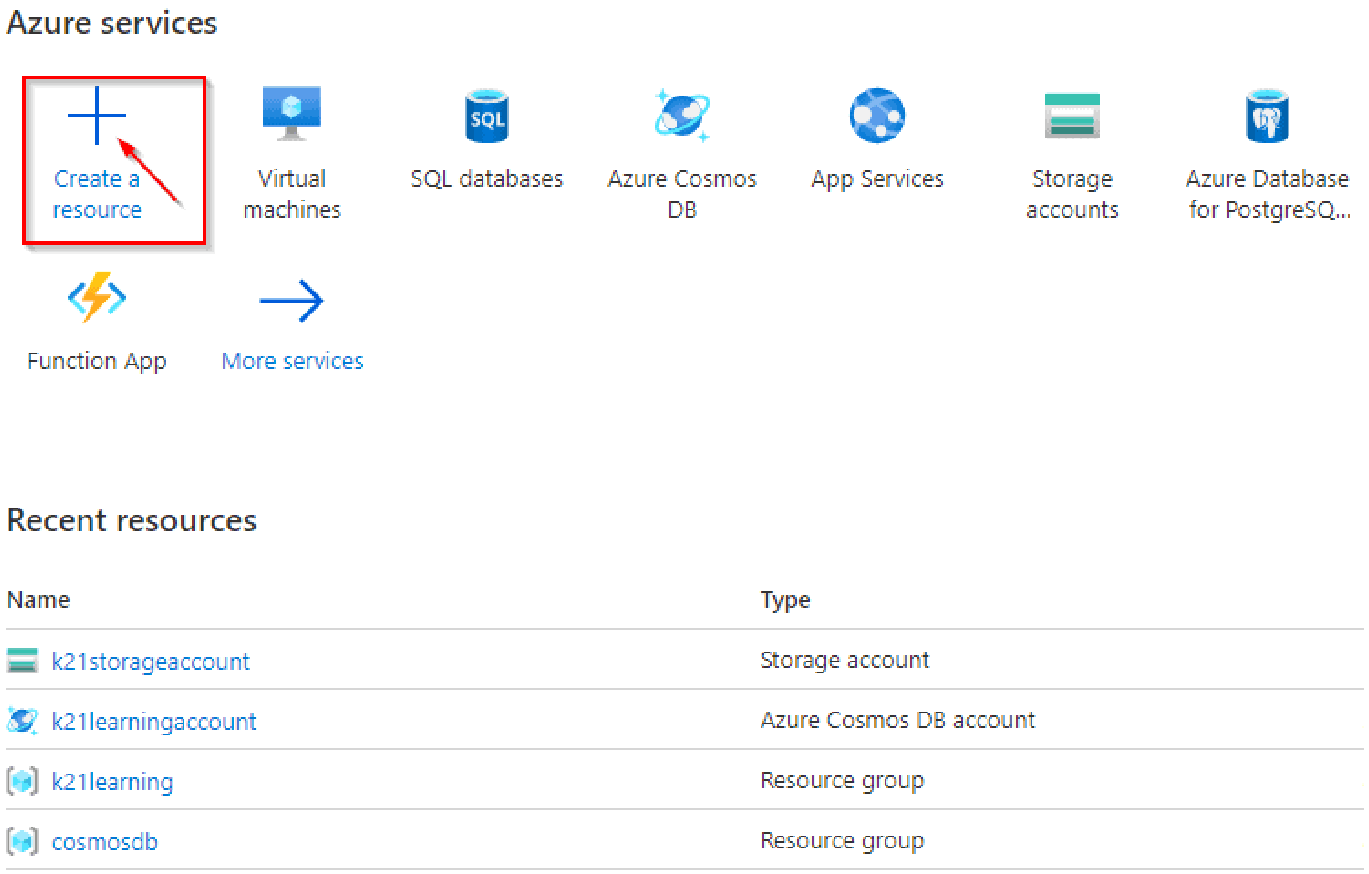

DATA SET
(e.g. table, file)
Consumes
Produces
ACTIVITY
(e.g. hive, stored proc., copy)
Is a logical grouping of
PIPELINE
(schedule, monitor, manage)
Represents a data item(s) stored in
Runs on
LINKED SERVICE
(e.g. SQL Server, Hadoop Cluster)

Also check: Overview of **Azure Stream Analytics**

How To Create An ADF

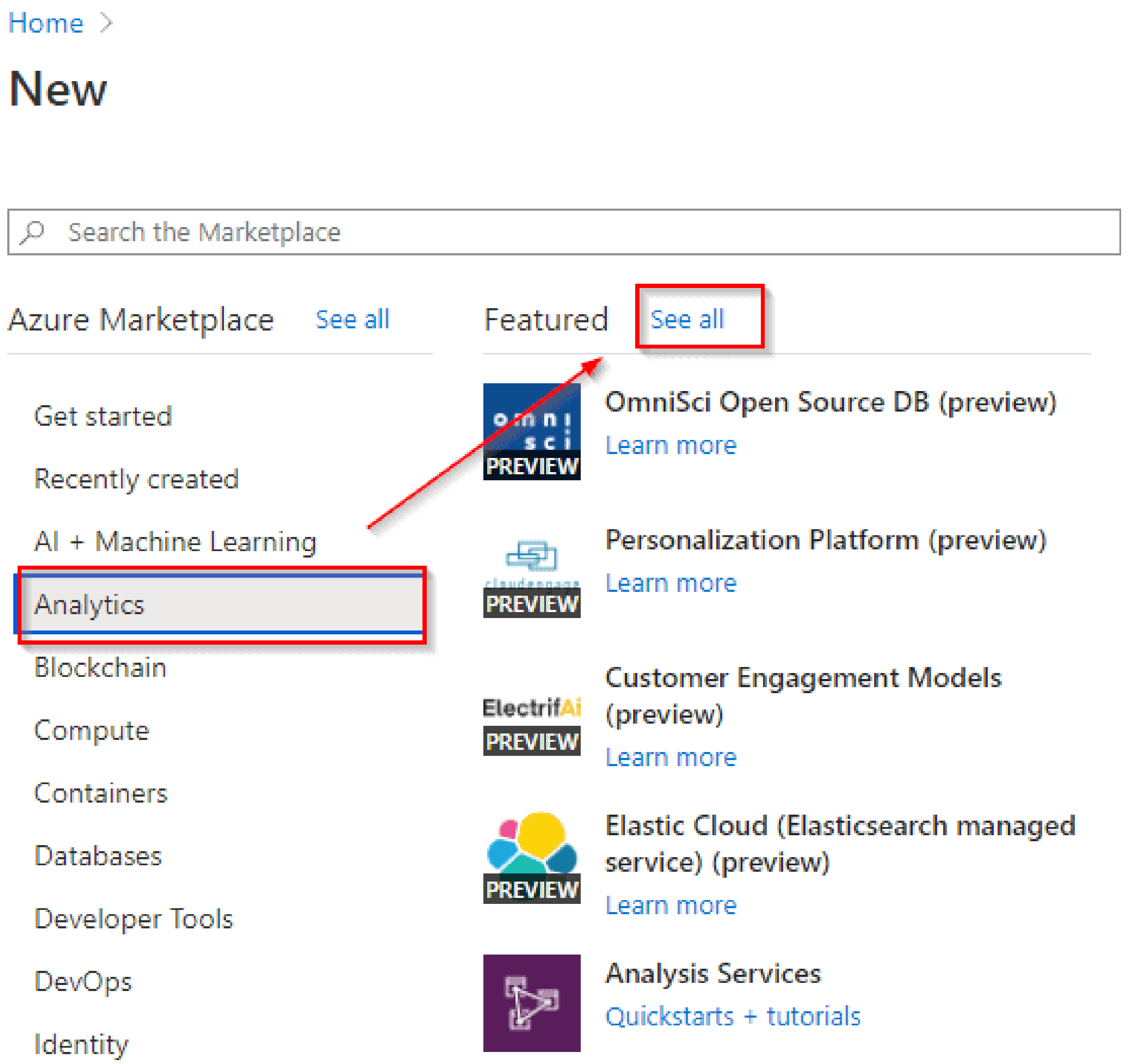
1) Go to the Azure portal.

2) From the portal menu, Click on Create a resource.



Name	Type
k21storageaccount	Storage account
k21learningaccount	Azure Cosmos DB account
k21learning	Resource group
cosmosdb	Resource group

3) Select Analytics, and then select see all.



Home >

New

Search the Marketplace

Azure Marketplace See all

Get started

Recently created

AI + Machine Learning

Analytics

Blockchain

Compute

Containers

Databases

Developer Tools

DevOps

Identity

Featured See all

OmniSci Open Source DB (preview)
Learn more

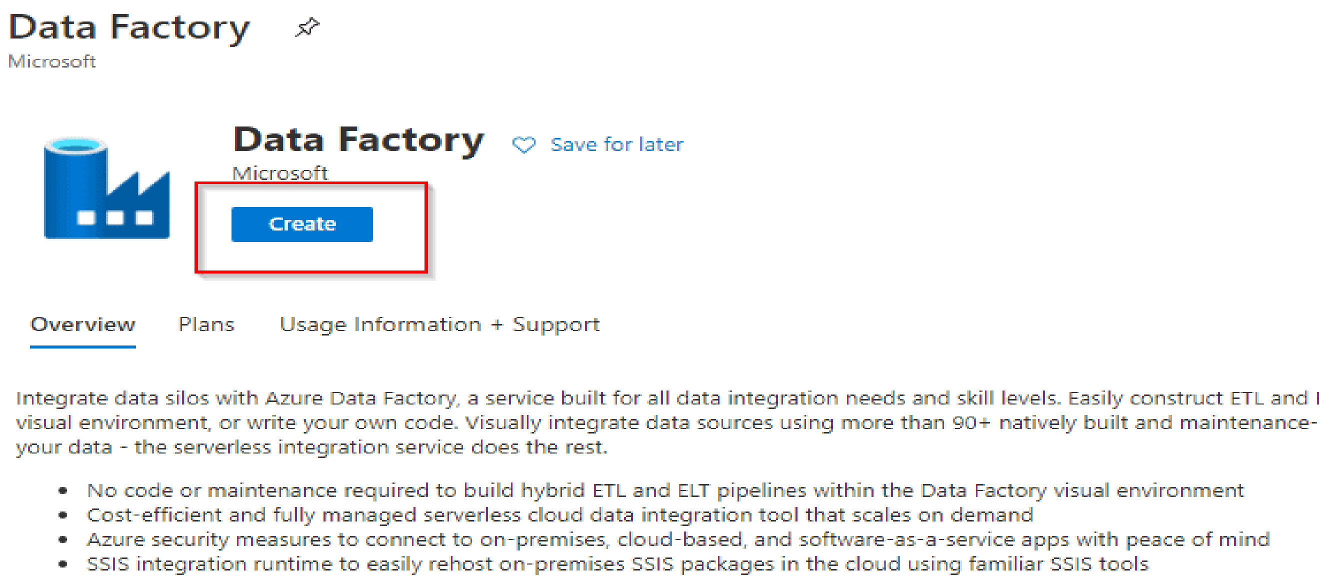
Personalization Platform (preview)
Learn more

Customer Engagement Models (preview)
Learn more

Elastic Cloud (Elasticsearch managed service) (preview)
Learn more

Analysis Services
Quickstarts + tutorials

4) Select Data Factory, and then select Create



Data Factory
Microsoft

Data Factory ♡ Save for later
Microsoft

Create

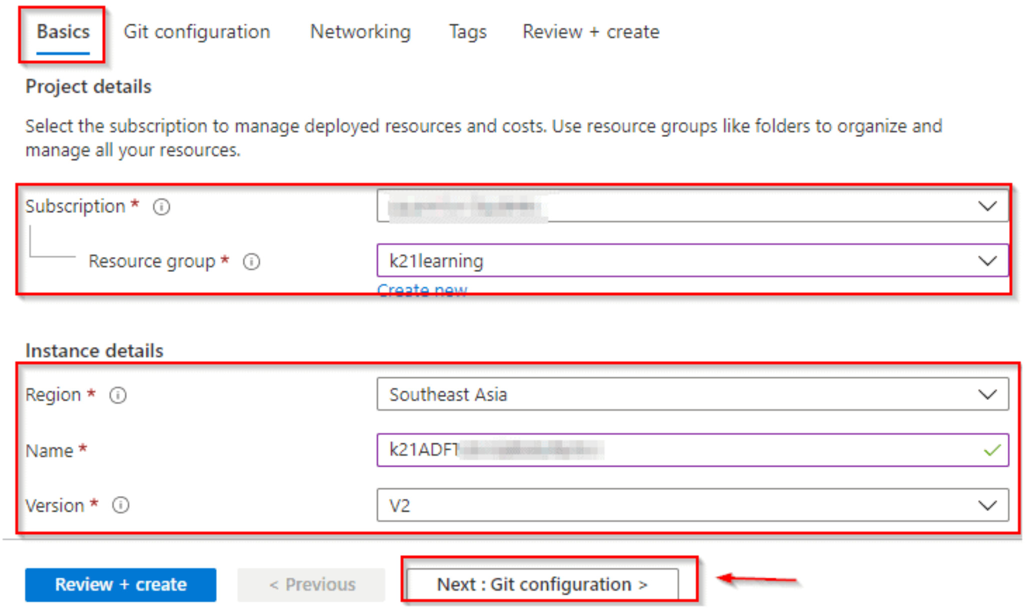
Overview Plans Usage Information + Support

Integrate data silos with Azure Data Factory, a service built for all data integration needs and skill levels. Easily construct ETL and I visual environment, or write your own code. Visually integrate data sources using more than 90+ natively built and maintenance- your data - the serverless integration service does the rest.

- No code or maintenance required to build hybrid ETL and ELT pipelines within the Data Factory visual environment
- Cost-efficient and fully managed serverless cloud data integration tool that scales on demand
- Azure security measures to connect to on-premises, cloud-based, and software-as-a-service apps with peace of mind
- SSIS integration runtime to easily rehost on-premises SSIS packages in the cloud using familiar SSIS tools

Check Out: How to create an **Azure load balancer**: step-by-step instruction for beginners.

5) On the Basics Details page, Enter the following details. Then Select Git Configuration.



Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * [redacted]

Resource group * k21learning

Instance details

Region * Southeast Asia

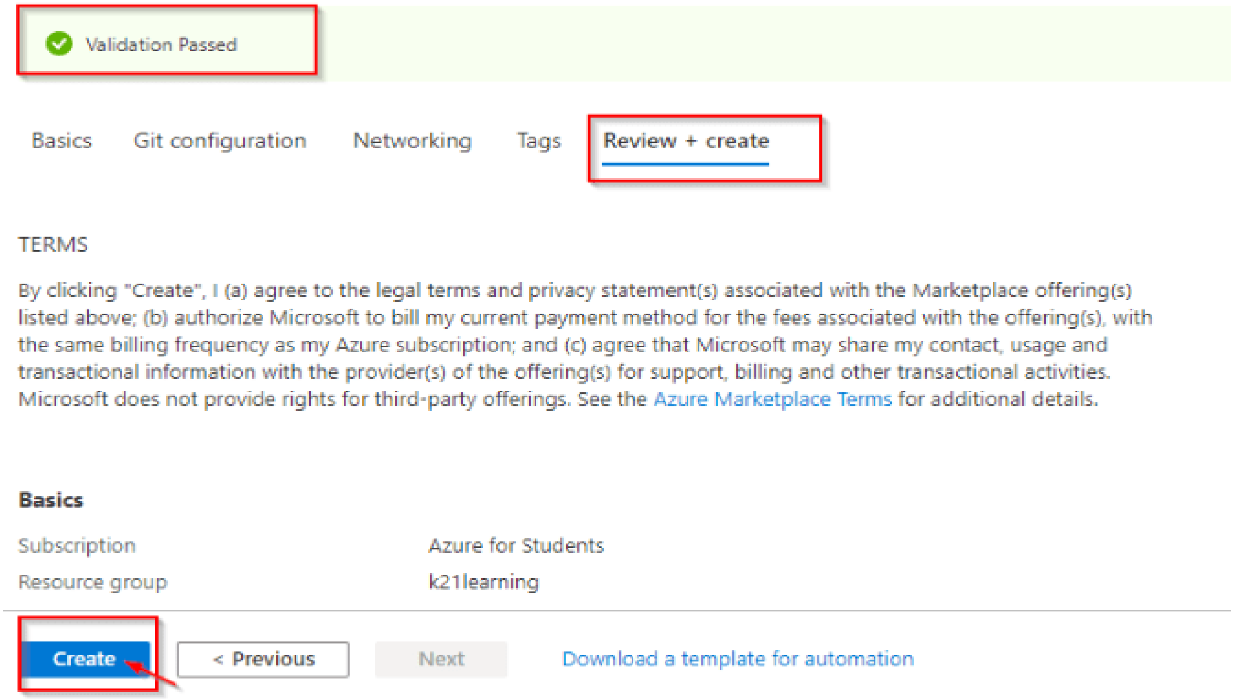
Name * k21ADF1[redacted]

Version * V2

6) On the Git configuration page, Select the Check the box, and then Go to Networking.

Also Check: Data Science VS Data Engineering, to know the major differences between them.

7) On the Networking page, don't change the default settings and click on Tags, and the Select Create.



Validation Passed

TERMS

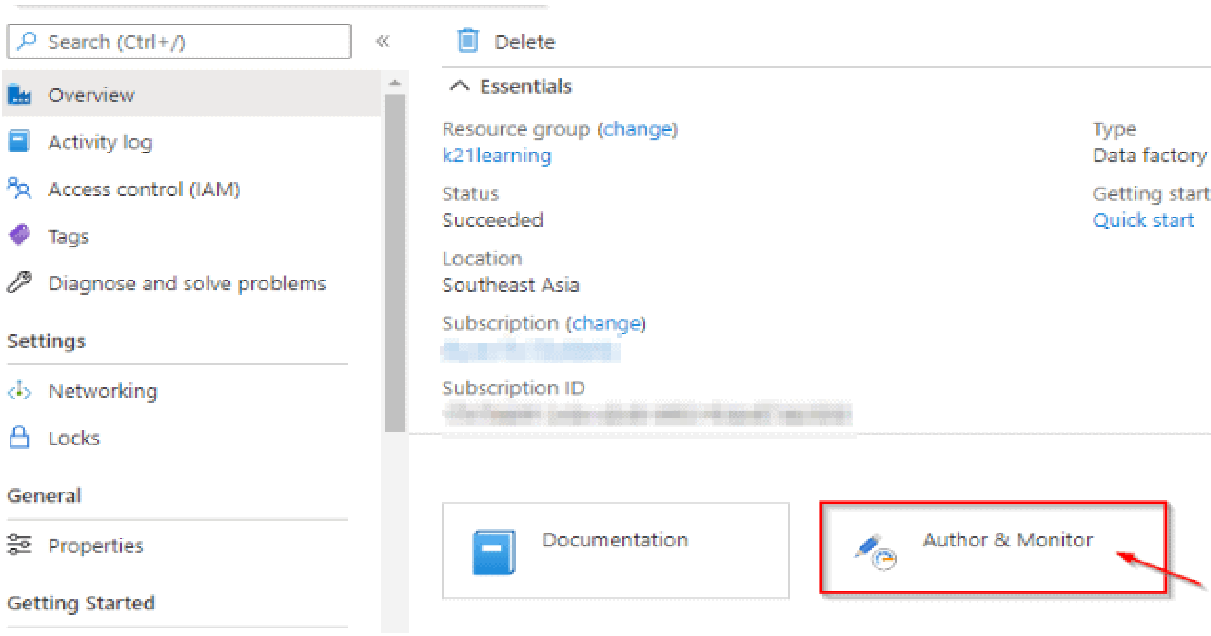
By clicking "Create", I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. See the Azure Marketplace Terms for additional details.

Basics

Subscription Azure for Students

Resource group k21learning

8) Select Go to resource, and then Select Author & Monitor to launch the Data Factory UI in a separate tab.



Resource group (change)
k21learning

Status
Succeeded

Location
Southeast Asia

Subscription (change)
[redacted]

Subscription ID
[redacted]

Type
Data factory (V2)

Getting started
Quick start

Frequently Asked Questions

Q: What is Azure Data Factory?

A: Azure Data Factory is a cloud-based data integration service provided by Microsoft. It allows you to create, schedule, and manage data pipelines that can move and transform data from various sources to different destinations.

Q: What are the key features of Azure Data Factory?

A: Azure Data Factory offers several key features, including data movement and transformation activities, data flow transformations, integration with other Azure services, data monitoring and management, and support for hybrid data integration.

Q: What are the benefits of using Azure Data Factory?

A: Some benefits of using Azure Data Factory include the ability to automate data pipelines, seamless integration with other Azure services, scalability to handle large data volumes, support for on-premises and cloud data sources, and comprehensive monitoring and logging capabilities.

Q: How does Azure Data Factory handle data movement?

A: Azure Data Factory uses data movement activities to efficiently and securely move data between various data sources and destinations. It supports a wide range of data sources, such as Azure Blob Storage, Azure Data Lake Storage, SQL Server, Oracle, and many others.

Q: What is the difference between Azure Data Factory and Azure Databricks?

A: While both Azure Data Factory and Azure Databricks are data integration and processing services, they serve different purposes. Azure Data Factory focuses on orchestrating and managing data pipelines, while Azure Databricks is a big data analytics and machine learning platform.

Q: Can Azure Data Factory be used for real-time data processing?

A: Yes, Azure Data Factory can be used for real-time data processing. It provides integration with Azure Event Hubs, which enables you to ingest and process streaming data in real time.

Q: How can I monitor and manage data pipelines in Azure Data Factory?

A: Azure Data Factory offers built-in monitoring and management capabilities. You can use Azure Monitor to track pipeline performance, set up alerts for failures or delays, and view detailed logs. Additionally, Azure Data Factory integrates with Azure Data Factory Analytics, which provides advanced monitoring and diagnostic features.

Q: Does Azure Data Factory support hybrid data integration?

A: Yes, Azure Data Factory supports hybrid data integration. It can connect to on-premises data sources using the Azure Data Gateway, which provides a secure and efficient way to transfer data between on-premises and cloud environments.

Q: How can I schedule and automate data pipelines in Azure Data Factory?

A: Azure Data Factory allows you to create schedules for data pipelines using triggers. You can define time-based or event-based triggers to automatically start and stop data pipeline runs.

Q: What security features are available in Azure Data Factory?

A: Azure Data Factory provides several security features, including integration with Azure Active Directory for authentication and authorization, encryption of data at rest and in transit, and role-based access control (RBAC) to manage access to data and pipelines. Please note that these FAQs are intended to provide general information about Azure Data Factory, and for more specific details, it is recommended to refer to the official Microsoft documentation or consult with Azure experts.