Azure Data Factory

Azure is Microsoft's comprehensive cloud computing platform and infrastructure for building, deploying, managing, and innovating applications and services through a global network of Microsoft-managed data centers. Azure provides an exceptionally extensive and continuously expanding set of cloud services, including compute, storage, networking, databases, analytics, AI, IoT, developer tools, mixed reality, security, management, monitoring, integration, governance, and more.

Highly Capable Compute

Azure provides extremely scalable on-demand computing resources like virtual machines, virtual machine scale sets, containers, serverless computing, batch processing, and service fabric. You can deploy both Windows Server and Linux virtual machines and scale them exceptionally to meet your needs. Azure offers sizes ranging from shared CPU cores and gigabytes of memory to tremendously large clusters with tens of thousands of dedicated cores and petabytes of memory. Azure virtual machines support Microsoft software like SQL Server, SharePoint, and Dynamics as well as open-source software like MySQL, PostgreSQL, MongoDB, and more. Azure Container Instances and Azure Kubernetes Service provide containers to package and run your applications at massive scale. Serverless computing options like Azure Functions and Logic Apps enable you to run code or workflows without managing infrastructure. Azure Batch provides a platform for running extraordinarily large-scale parallel and high-performance computing workloads efficiently in the cloud.

Massively Scalable Storage

Azure provides extraordinarily scalable storage options like Blob storage for object storage, File storage for file shares, Disk storage for block storage, and Data Lake Storage for big data analytics solutions. You can store and retrieve exceptionally large amounts of data in Azure. Storage options include hot, cool, and archive access tiers based on how frequently you need to access your data. Azure Storage is designed to be massively scalable to support the storage needs of today's most demanding

applications. You can use Azure Storage to store unstructured data, files, messages, backups, and disaster recovery at tremendous scale. Azure Files provides file shares in the cloud that are accessible via the Server Message Block (SMB) protocol. Azure Data Lake Storage is optimized for big data analytics workloads at massive scale. Azure Blob Storage provides object storage for cloud-based workloads.

Robust and Global Networking

Azure provides an exceptionally robust set of networking capabilities like virtual networks, subnets, network interfaces, load balancers, application gateways, VPN gateways, DNS zones, traffic manager, Azure Firewall, Azure Front Door, Azure Bastion, and more. You can connect your on-premises networks to Azure using site-to-site VPNs or Azure ExpressRoute. Azure's networking services help you securely extend your on-premises networks into the cloud and enable your Azure resources to connect to each other. Azure Load Balancer and Application Gateway provide load balancing and traffic routing across your Azure resources. You can configure virtual private networks, private link, network security groups, and service endpoints to control access to Azure resources. Azure DNS provides ultra-fast, reliable DNS for your Azure resources. Azure Front Door provides global HTTP load balancing and acceleration for your applications. Azure Bastion provides secure and seamless RDP and SSH access to your virtual machines directly through the Azure Portal.

Databases for All Your Needs

Azure provides database services like Azure SQL Database, Azure Database for MySQL, Azure Database for PostgreSQL, Azure Cosmos DB, Azure Database for MariaDB, Azure Cache for Redis, Azure Database Migration Service, Azure SQL Managed Instance, and Azure Database for PostgreSQL—Hyperscale (Citus). Azure SQL Database is a relational database service based on SQL Server. Azure Database for MySQL, PostgreSQL, and MariaDB are managed database services for open-source relational databases. Azure Cosmos DB is a globally distributed multi-model database service that supports document, graph, and key-value data models. Azure Cache for Redis provides a managed Redis cache to improve the performance and scalability of your applications. The Azure Database Migration Service can migrate your on-premises SQL Server, Oracle, and MySQL databases to Azure. Azure SQL Managed Instance provides a fully managed SQL Server instance in Azure. Azure Database for PostgreSQL—Hyperscale (Citus) provides a distributed database built on PostgreSQL.

Analytics, AI, and IoT at Massive Scale

Azure provides analytics services like Azure Databricks, Azure Data Lake Analytics, Azure Machine Learning, Power BI, Azure Stream Analytics, Azure Time Series Insights, HDInsight, and Azure Data Explorer. You can build exceptionally advanced machine learning and AI solutions with Azure Machine Learning. Azure IoT Hub enables you to connect, monitor, and manage billions of Internet of Things assets. Azure Stream Analytics provides real-time analytics on fast moving streams of data from applications, websites, sensors, social media, and more at tremendous scale. Power BI is a suite of business analytics tools to analyze data and share insights. Azure Databricks provides a collaborative Apache Spark-based analytics platform. HDInsight provisions managed Hadoop, Spark, Kafka, and HBase clusters in Azure. Azure Time Series Insights provides a platform for industrial IoT applications. Azure Data Explorer is a fast and highly scalable data exploration service for log and telemetry data.

Integrate and Automate Everything

Azure provides integration services like Azure Event Grid, Service Bus, Logic Apps, API Management, Azure Data Factory, and Azure DevOps Services. Azure Event Grid provides event routing service that enables you to manage events from Azure services and your own applications. Azure Service Bus provides messaging services to connect applications and services. Azure Logic Apps provides a visual designer to model and automate workflows. Azure API Management helps you publish APIs to developers, partners, and employees securely and at scale. Azure Data Factory is a cloud-based data integration service for orchestrating and automating data movement and transformation. Azure DevOps Services provides development collaboration tools including high-performance pipelines, free private Git repositories, and agile tools.

Enterprise-Grade Security

Azure provides a set of security services like Azure Active Directory, Azure Key Vault, Azure Information Protection, Azure Security Center, Azure Sentinel, Azure Firewall, and Azure Dedicated HSM. Azure Active Directory provides identity and access management for your cloud applications. Azure Key Vault helps you safeguard cryptographic keys and secrets. Azure Information Protection helps you classify and label data. Azure Security Center provides security monitoring and policy management across your Azure resources. Azure Sentinel provides a scalable, cloud-native, security information event management (SIEM) and security orchestration automated response

(SOAR) solution. Azure Firewall provides a managed, stateful firewall as a service with built-in high availability and unrestricted cloud scalability. Azure Dedicated HSM provides hardware security modules (HSMs) as a service to help secure your keys and secrets.

Manage and Monitor Everything

Azure provides services for managing and monitoring your solutions like Azure Resource Manager, Azure Monitor, Log Analytics, Application Insights, Azure Automation, Azure Policy, Azure Blueprints, and Azure Advisor. Azure Resource Manager enables you to deploy, manage, and monitor all your Azure resources. Azure Monitor provides a comprehensive solution for collecting, analyzing, and acting on telemetry from your cloud and on-premises environments. Log Analytics provides a log aggregation and analytics solution for your Azure and on-premises resources. Application Insights provides application performance management for your live services. Azure Automation provides a way for you to automate the deployment, configuration, and management of your Azure resources. Azure Policy provides governance and compliance for your Azure resources. Azure Blueprints provides a way to define repeatable Azure deployments. Azure Advisor provides recommendations to help you follow best practices and optimize your Azure deployments.

Govern Your Resources

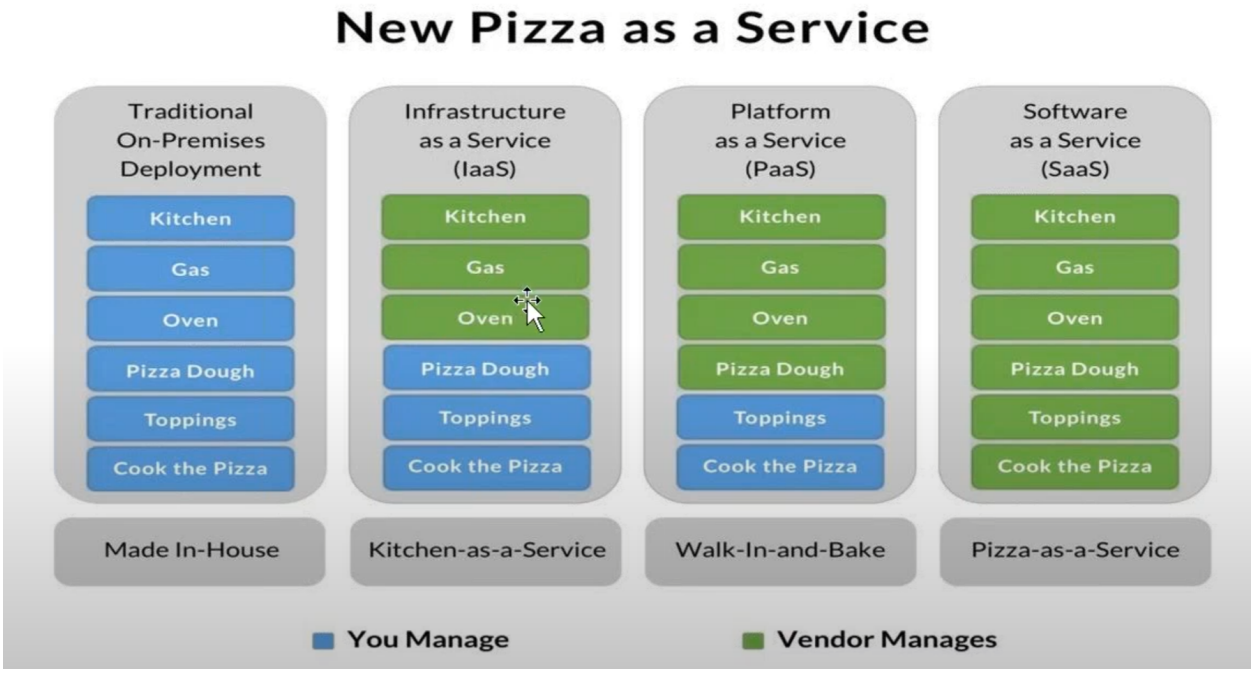
Azure provides governance services like Azure Policy, Azure Blueprints, Azure Resource Locks, Azure Management Groups, and Azure Role-Based Access Control (RBAC). Azure Policy provides governance and compliance for your Azure resources through policies that enforce different rules and effects. Azure Blueprints provides a way to define repeatable Azure deployments through blueprints that contain Azure resources and artifacts. Azure Resource Locks provide a mechanism to lock Azure resources to prevent accidental deletion or modification. Azure Management Groups provide a level of scope above subscriptions to organize multiple subscriptions. Azure RBAC provides granular access management for Azure resources.

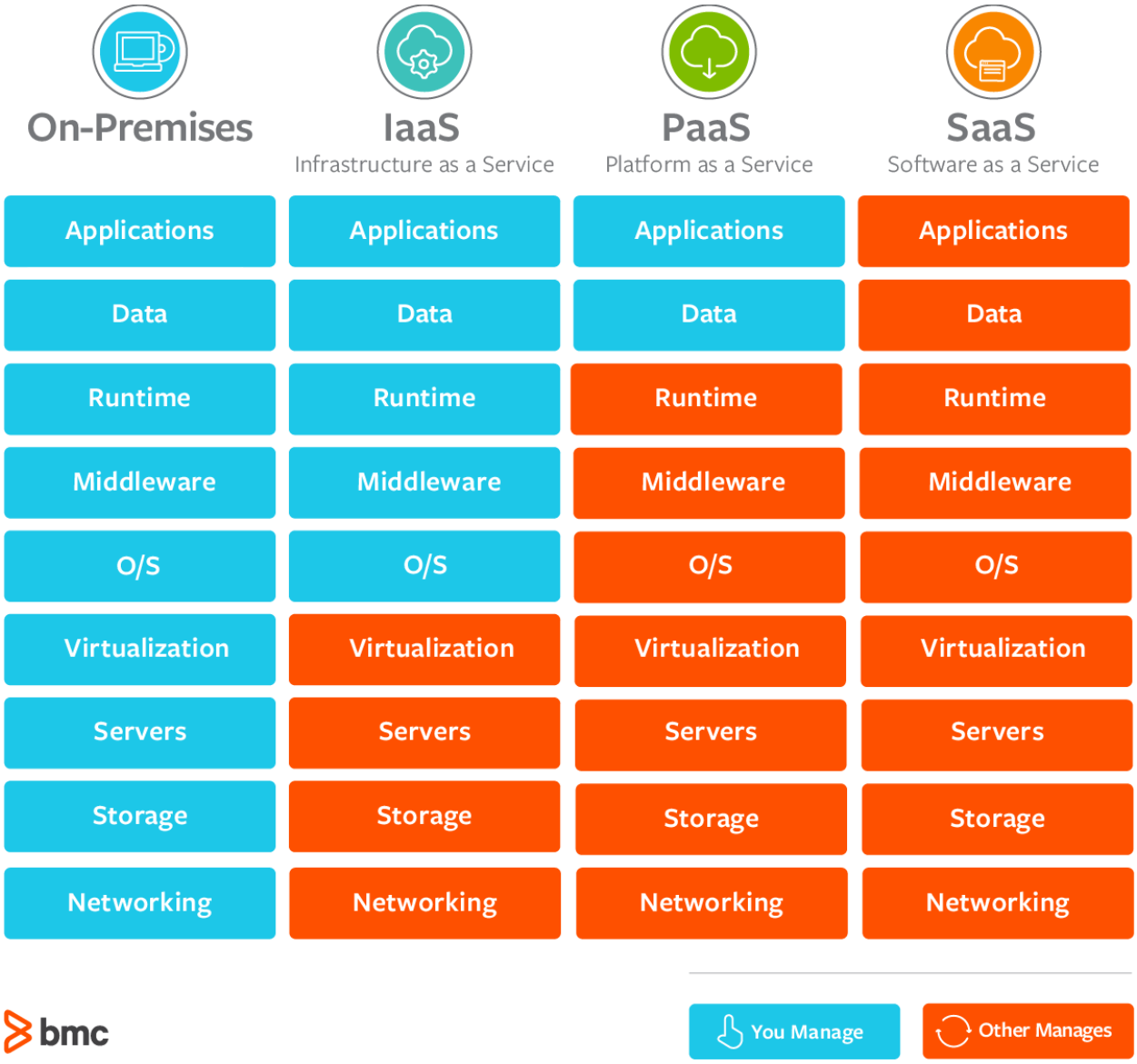
Conclusion

In conclusion, Azure provides an extraordinarily comprehensive and continuously expanding set of cloud services that can help you build and manage your applications. From computing resources to storage, networking, databases, analytics, AI, IoT,

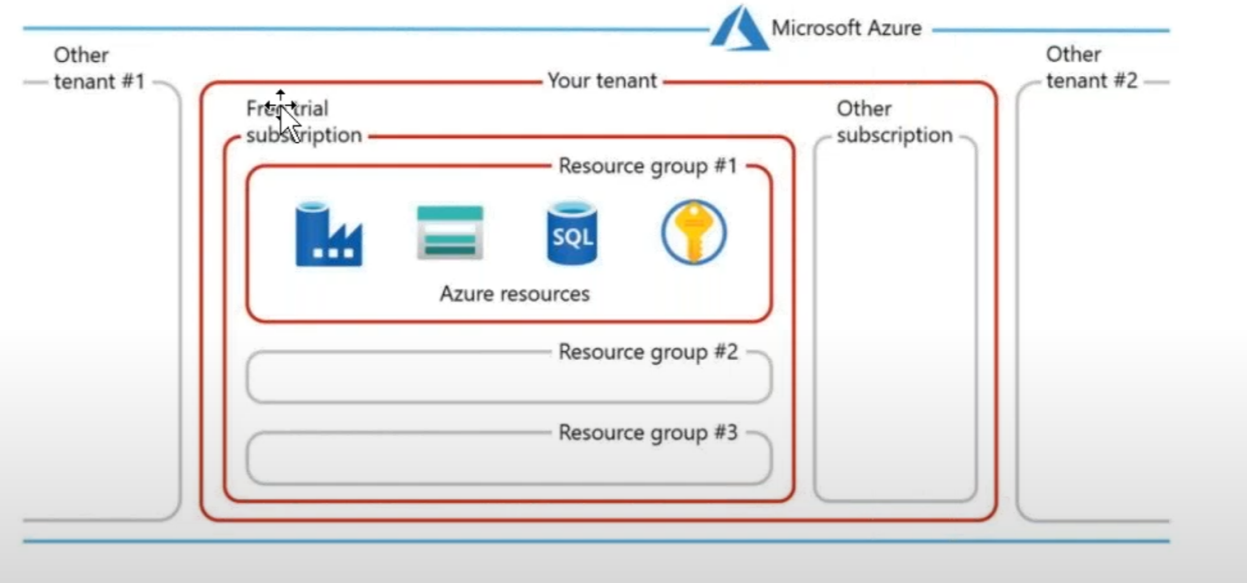
security, management, monitoring, integration, governance, and more, Azure has become an industry leader in enterprise cloud computing. Azure's services are designed to simplify infrastructure management, increase developer productivity, and accelerate the development of new solutions. With Azure, you can build and manage applications anywhere using the tools, languages, and frameworks you want. Azure is an open and flexible platform with a thriving partner ecosystem, making it the platform of choice for your cloud solutions. Azure provides innovative services, a commitment to security and privacy, and the ability to meet compliance standards like ISO, HIPAA, FedRAMP, and more. Azure enables you to achieve more by helping your organization be more agile, reduce costs, and focus on innovation.

Iaas vs Paas vs Saas





Resource Groups

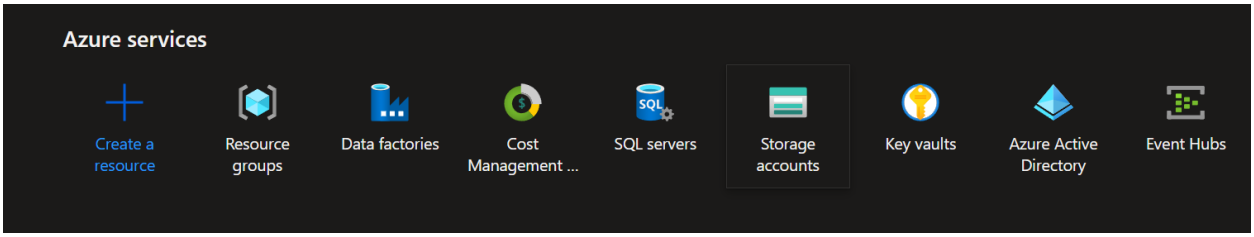


Tenant = Client / Project

In Azure, a resource group is a logical container for grouping and managing related Azure resources. Azure resources such as virtual machines, storage accounts, databases, and virtual networks are typically deployed together to support an application or a solution. By organizing these resources in a resource group, you can manage them together and apply common policies and access control rules.

A resource group can be thought of as a boundary for managing and monitoring the resources in your Azure subscription. You can create, update, and delete all resources in a resource group together. You can also monitor the health and performance of resources in a resource group, view their usage and costs, and set up alerts and notifications for any changes or issues.

Resource groups also help with billing and cost management, as all resources in a group are billed together based on their usage. You can also assign tags to resources in a resource group to categorize and track their costs and usage.



Click on resource groups and then select the subscription, provide a username and select a location. Then click on create and resource group will be created.

Services offered by Azure

- Compute services

1. Virtual Machines (VMs): This is a compute service that allows you to run Windows or Linux-based virtual machines in the cloud. You can choose from a wide range of pre-configured VM images or create your own custom image. You can also choose the size and configuration of your VMs based on your workload requirements.
2. App Service: This is a platform-as-a-service (PaaS) offering that allows you to quickly build, deploy, and scale web and mobile applications. App Service supports a range of programming languages and frameworks, including .NET, Java, Node.js, PHP, Python, and Ruby.
3. Functions: This is a serverless compute service that allows you to run code in response to events, such as a new file being uploaded to Azure Storage or a message being posted to a queue. Functions scales automatically and charges you only for the time your code runs.
4. Kubernetes Service (AKS): This is a managed Kubernetes service that allows you to deploy and manage containerized applications at scale. AKS automates many of the manual processes involved in deploying and managing Kubernetes clusters.
5. Batch: This is a platform for running large-scale parallel and high-performance computing (HPC) workloads. Batch automatically scales compute resources based on workload demands, and supports a range of HPC workloads, including scientific simulations, rendering, and deep learning.

- Storage Services

1. Blob Storage: Blob storage is used to store unstructured data such as text, images, and videos. It provides a highly scalable and cost-effective way to store and access large amounts of data.
2. File Storage: File storage is used to store and share files within and across organizations. It is similar to a network file share and is used for applications that require a file system interface.

3. Table Storage: Table storage is used to store structured NoSQL data, such as JSON documents. It provides a key-value store with a schema-less design.
4. Queue Storage: Queue storage is used to store messages that need to be processed asynchronously. It provides a reliable and scalable messaging solution for decoupling applications.
5. Disk Storage: Disk storage is used to store persistent virtual machine disks. It provides high-performance, reliable, and scalable storage for virtual machines running in Azure.

- Networking Services

1. Virtual Network (VNet): This is a logically isolated network in Azure that enables you to create and manage your own IP address space, subnets, routing, and security policies. You can use VNets to connect your Azure resources, such as virtual machines, to each other, and to your on-premises network via VPN or Azure ExpressRoute.
2. Azure Load Balancer: This is a highly available, scalable, and low-latency load balancing service that distributes incoming traffic across multiple backend resources, such as virtual machines, virtual machine scale sets, or availability sets. You can use Azure Load Balancer to improve the availability, performance, and scalability of your applications.
3. Azure Application Gateway: This is a web traffic load balancer that provides advanced application delivery features, such as SSL offloading, URL-based routing, and web application firewall (WAF). You can use Azure Application Gateway to optimize the performance and security of your web applications.
4. Azure Traffic Manager: This is a global DNS-based traffic management service that allows you to distribute user traffic across multiple endpoints, such as Azure regions, external endpoints, or on-premises datacenters. You can use Azure Traffic Manager to improve the availability and performance of your applications for global users.
5. Azure VPN Gateway: This is a VPN appliance service that provides secure connectivity between your on-premises network and your Azure VNet. You can use Azure VPN Gateway to establish site-to-site VPN or point-to-site VPN connections with your on-premises network.

6. Azure ExpressRoute: This is a dedicated private network connection service that provides high-bandwidth, low-latency, and reliable connectivity between your on-premises network and your Azure VNet. You can use Azure ExpressRoute to extend your on-premises network into Azure, or to build hybrid applications that require low-latency connectivity to Azure services.

- App Hosting Services

1. Azure App Service: This is a fully managed platform for hosting web apps, mobile app backends, and RESTful APIs. You can deploy apps written in .NET, Node.js, Java, PHP, or Python, and integrate them with other Azure services such as Azure Functions, Azure Cosmos DB, and Azure DevOps.
2. Azure Kubernetes Service (AKS): This is a fully managed Kubernetes service that lets you deploy and manage containerized applications with ease. You can deploy apps to AKS using Docker containers, and scale them up or down based on demand.
3. Azure Functions: This is a serverless compute service that lets you run event-driven code in the cloud. You can write your code in C#, Java, JavaScript, Python, or PowerShell, and trigger it based on events from other Azure services or external sources.
4. Azure Logic Apps: This is a workflow automation service that lets you create and run workflows to integrate your apps and services. You can use pre-built connectors to connect to over 200 services, including Azure services, SaaS apps, and on-premises systems.
5. Azure Service Fabric: This is a distributed systems platform that lets you build and deploy microservices-based applications. You can choose from various programming models, including .NET, Java, and Node.js, and deploy your services to clusters that automatically handle scaling and fault tolerance.

- AI Services

1. Azure Cognitive Services: This is a collection of pre-built AI services that developers can use to add intelligent features to their applications. These services include speech recognition, language understanding, image and video analysis, and more.
2. Azure Machine Learning: This is a cloud-based platform that enables developers to build, train, and deploy machine learning models at scale. It includes a range of

tools and frameworks that can be used to build custom models for specific use cases.

3. Azure Bot Service: This is a platform for building and deploying intelligent bots that can interact with users through a variety of channels, including text messages, chat, and voice.
4. Azure Databricks: This is a fast, easy, and collaborative Apache Spark-based analytics platform that enables data scientists to build and train machine learning models at scale.
5. Azure Cognitive Search: This is a fully managed search service that enables developers to add search capabilities to their applications. It uses advanced AI algorithms to deliver relevant search results based on user queries.

- Integration Services

1. Azure Logic Apps - A service that allows you to build workflows and automate processes using a drag-and-drop visual designer. It supports hundreds of connectors for integrating with various services and applications.
2. Azure Service Bus - A messaging service that enables reliable and scalable communication between applications and services. It supports messaging patterns such as publish/subscribe, queues, and topics.
3. Azure Event Grid - A fully-managed event routing service that simplifies the development of event-driven applications. It supports events from various sources, including Azure services and custom applications.
4. Azure API Management - A service that enables you to publish, secure, and manage APIs. It provides features such as authentication, authorization, rate limiting, and analytics.
5. Azure Data Factory - A cloud-based data integration service that enables you to create, schedule, and orchestrate data workflows. It supports data movement, transformation, and integration across various data sources and destinations.

- Security Services

1. Azure Security Center: Provides a unified view of security across your Azure resources, detects and responds to threats, and helps you comply with industry standards and regulations.

2. Azure Active Directory: Provides identity and access management for Azure and other Microsoft services, including single sign-on, multifactor authentication, and conditional access policies.
3. Azure Firewall: Provides a network security service to protect your Azure Virtual Network resources.
4. Azure DDoS Protection: Helps protect your applications and services from distributed denial-of-service (DDoS) attacks.
5. Azure Information Protection: Helps you classify, label, and protect sensitive data using encryption, access policies, and other security controls.
6. Azure Key Vault: Provides a secure way to store and manage cryptographic keys, secrets, and certificates used by your applications and services.
7. Azure Advanced Threat Protection: Helps detect and investigate advanced attacks on your on-premises and cloud resources.
8. Azure Sentinel: Provides a cloud-native security information and event management (SIEM) service to analyze security data across your Azure resources and other sources.

Benefit of using Azure is “Only pay for what you use” & “Services are elastic as they can grow on demand”

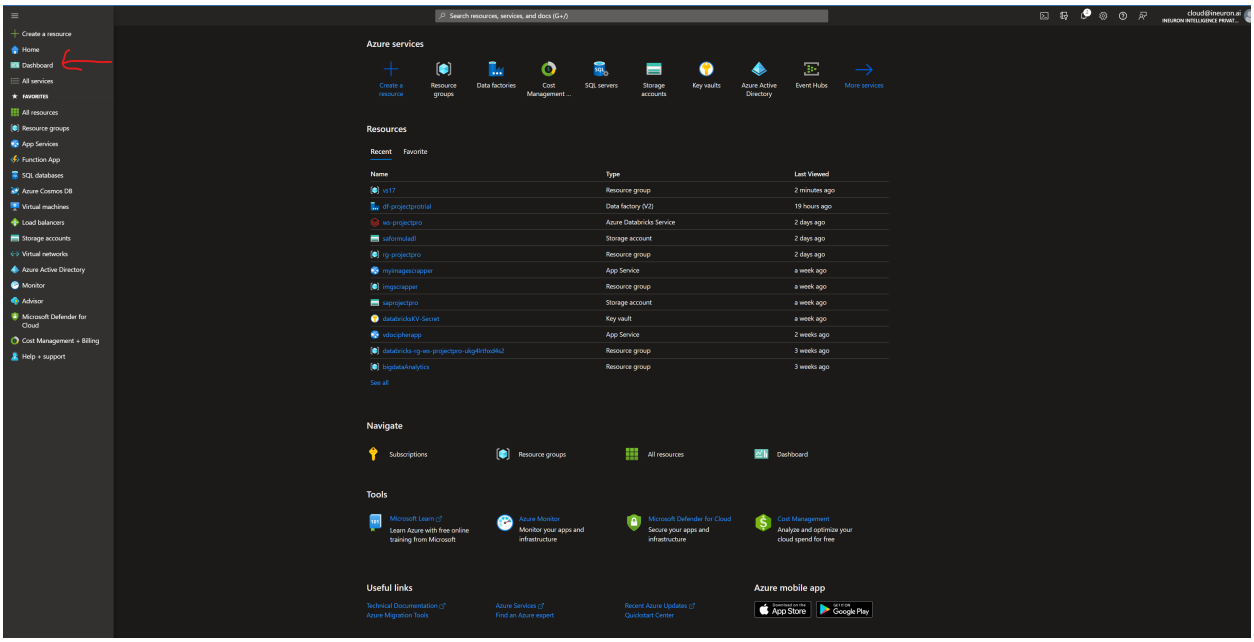
Creating personal dashboard

Creating a dashboard in Azure allows you to monitor and manage your Azure resources in a centralized and customizable way. Here are some of the benefits and uses of creating a dashboard in Azure:

1. Visualize your resources: A dashboard can display information about your Azure resources in a visually appealing and easy-to-understand format. You can create charts, graphs, tables, and other visualizations that show the health, performance, and usage of your resources.
2. Customization: You can customize your dashboard to show the specific information you need, and arrange it in a way that makes sense to you. You can choose which metrics to display, how they are presented, and how often they are updated.

3. Monitoring: A dashboard allows you to monitor the health and performance of your Azure resources in real-time. You can set up alerts and notifications for specific metrics or conditions, and be notified when there is an issue that requires your attention.
4. Collaboration: You can share your dashboard with other users in your organization, allowing them to view and monitor the same resources. This can facilitate collaboration and improve communication across teams.
5. Cost management: A dashboard can help you monitor your Azure costs and usage, and identify areas where you can optimize your spending. You can create charts and graphs that show your spending trends, and drill down into specific resources to see their costs and usage.

Overall, creating a dashboard in Azure provides a centralized and customizable view of your Azure resources, allowing you to monitor, manage, and optimize your resources more effectively.



SDK and Tools provided by Azure

https://azure.microsoft.com/en-in/downloads/

Azure Data Factory

Azure data factory is Azure's cloud ETL service for scale-out serverless data integration and data transformation. You can also lift and shift existing SSIS packages to Azure and run them with full compatibility in Azure data factory.

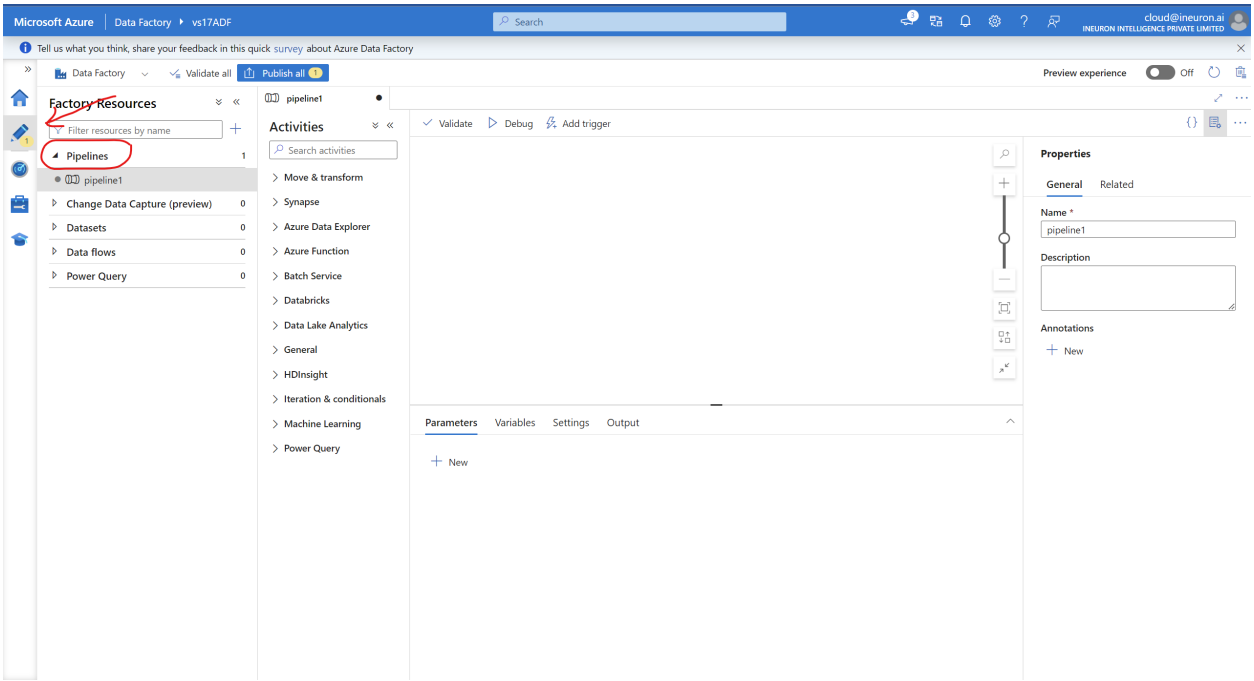
It is a cloud based data integration service that allow us to create data driven workflows for orchestrating data movement and transforming data at scale.

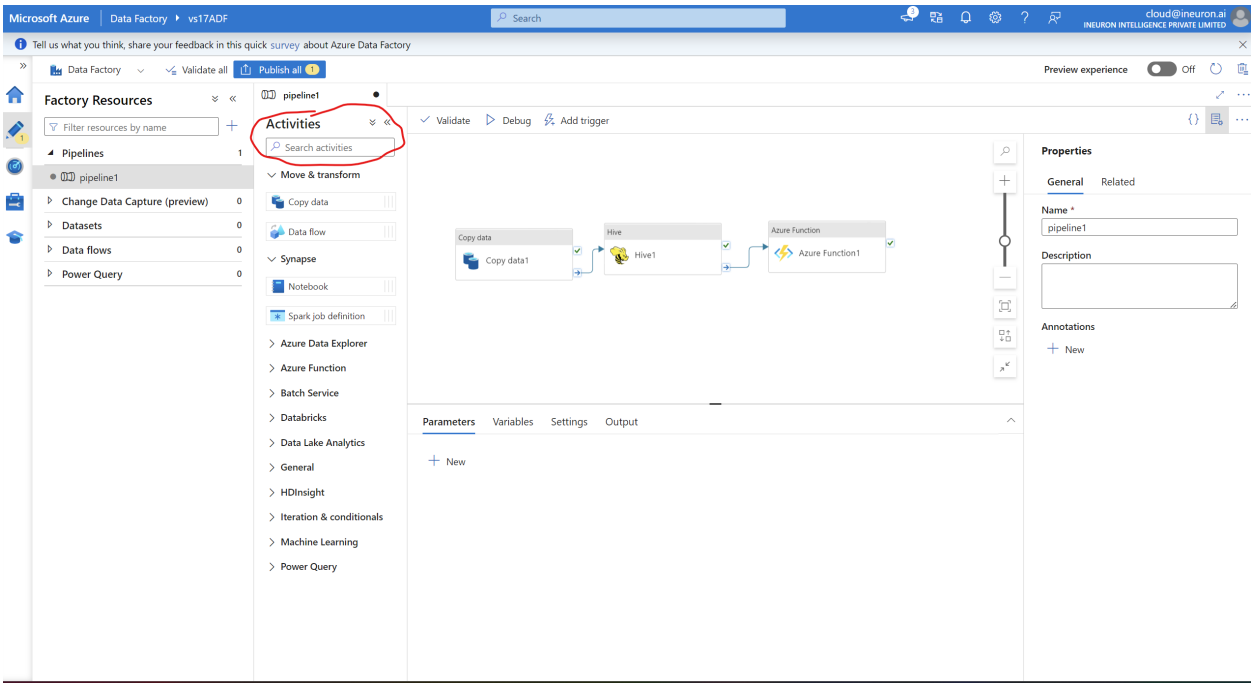
Why ADF?

Big data requires service that can orchestrate and operationalize processes to refine these enormous stores of raw data into actionable business insights. ADF is a managed cloud service that's built for complex hybrid ETL or ELT or data ingestion projects.

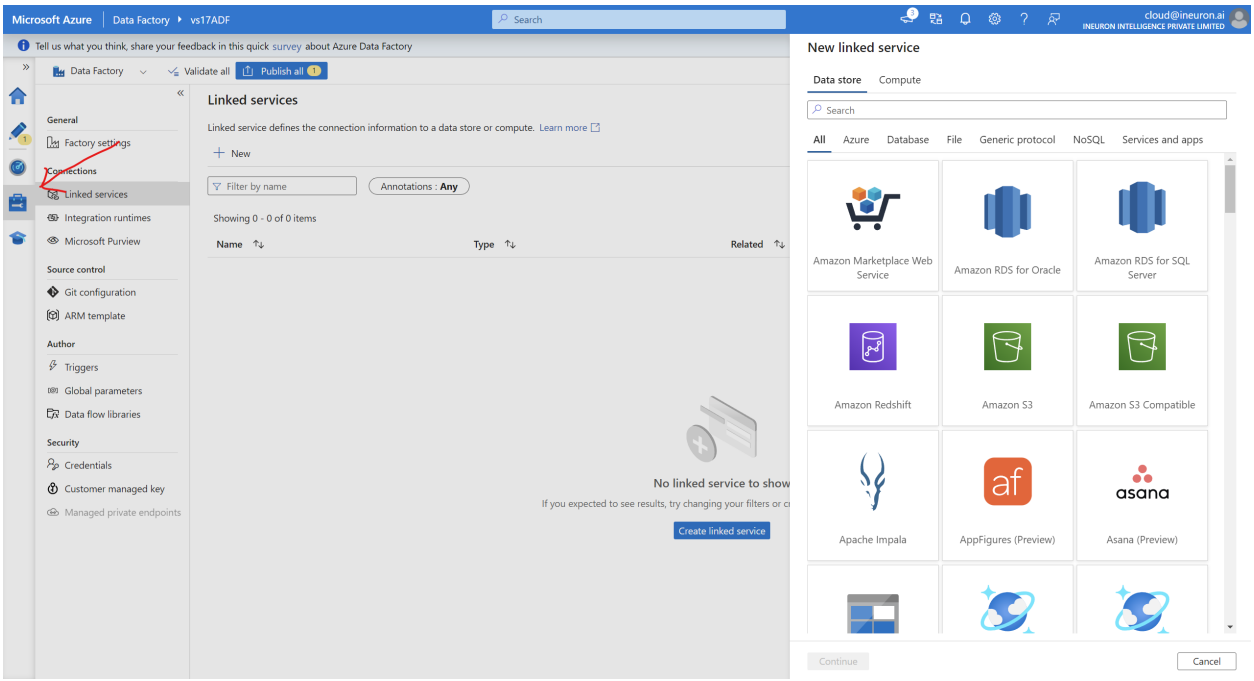
Top level components of ADF

- Pipeline → ADF might have one or more pipelines. A pipeline is logical grouping activities that performs a unit of work. For e.g., a pipeline can contain a group of activities that ingests data from Azure Blob and then runs a Hive query on an HDInsight cluster to partition the data.



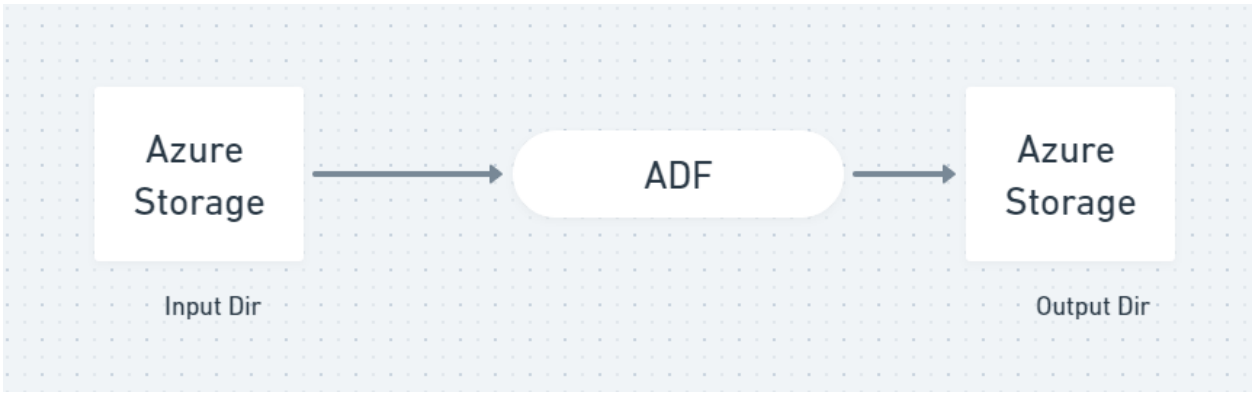


- Activity → Activity represents a processing step in a pipeline. For e.g., you might use a copy activity to copy data from one data store to another data store. Running the Hive query is another activity.
- Linked Services → Linked services are much like connection strings, which define the connection information that's needed for Data Factory to connect to external reoureces.

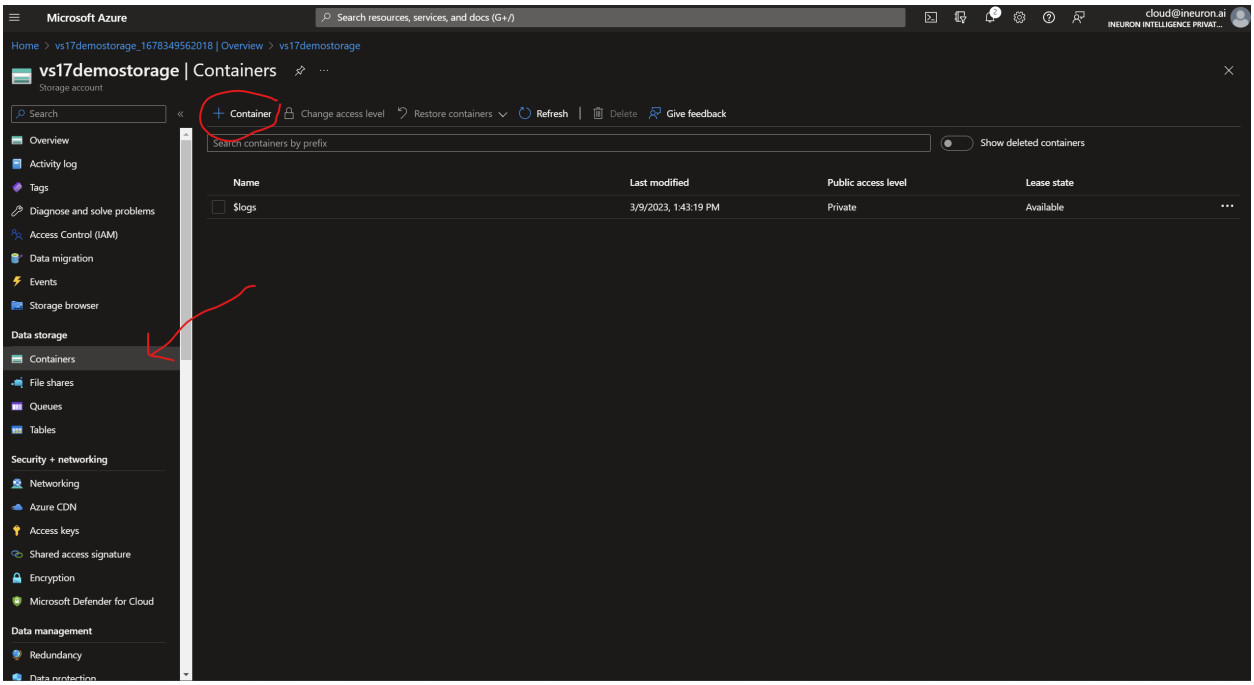


- Datasets → Datasets represent data structures within the data stores, which simply point to or reference the data we want to use in our activities. For e.g., an Azure storage linked service specifies a connection string to connect to the Azure Storage account. Additionally, an Azure blob dataset specifies the blob container and the folder that contains the data.
- Triggers → Triggers determines when a pipeline needs to run.

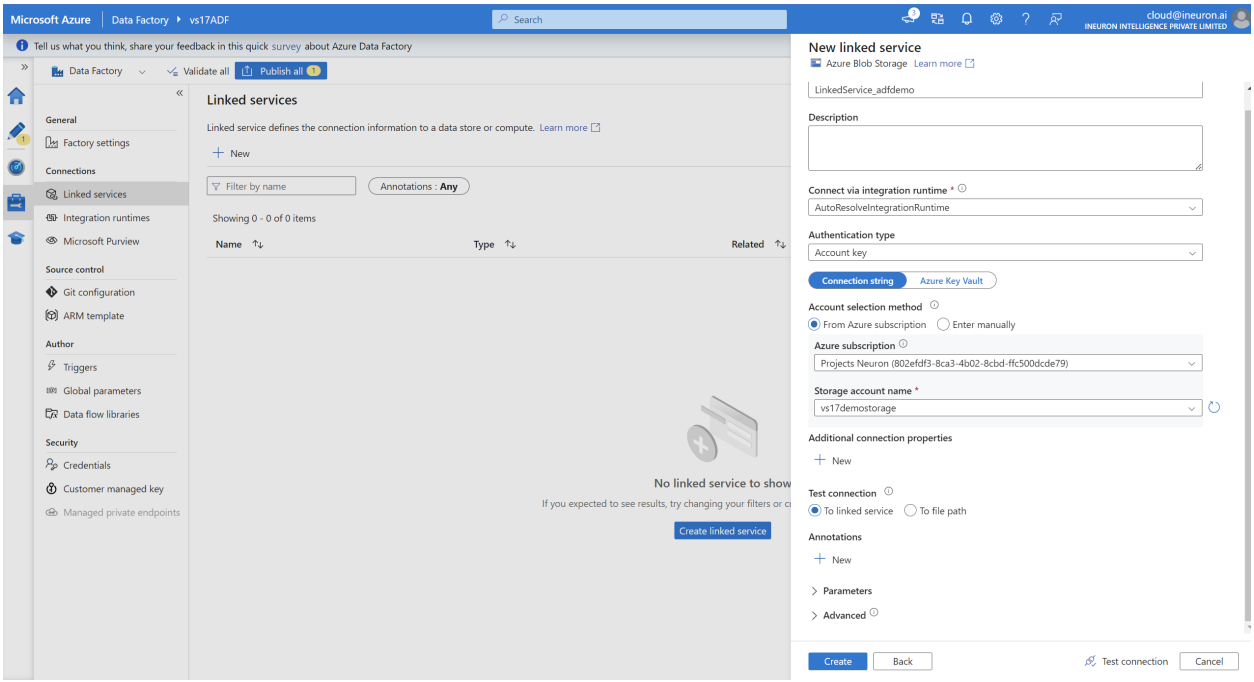
Create our first data factory



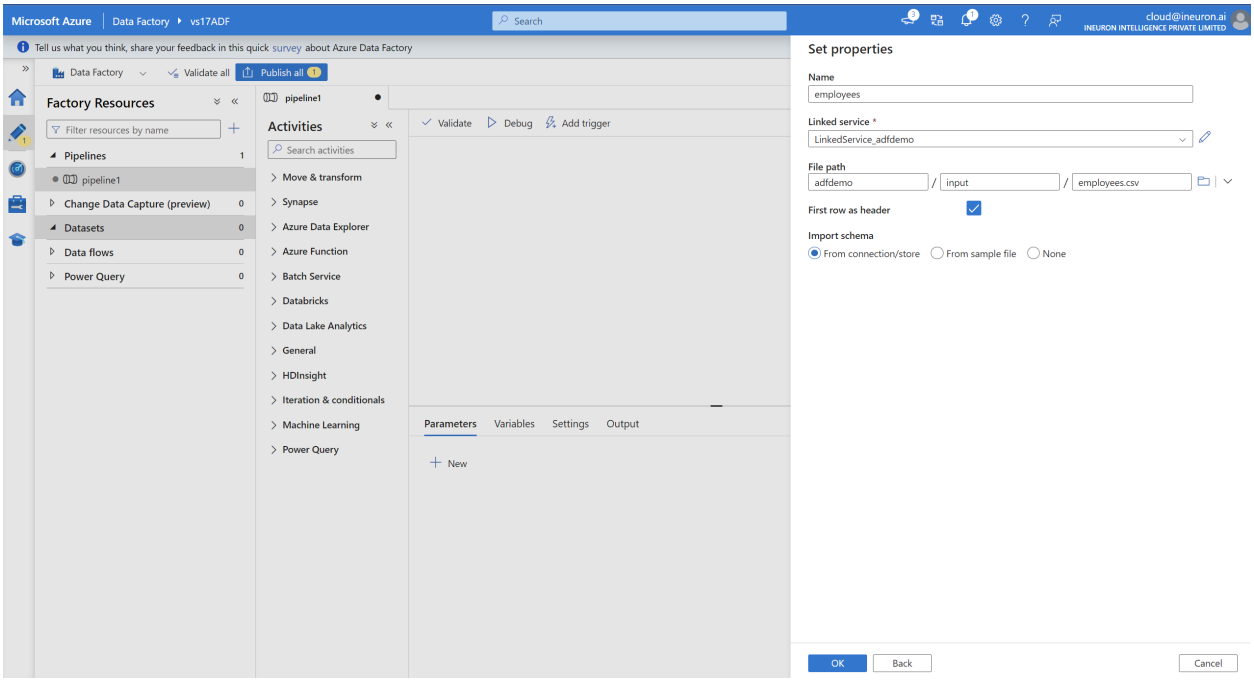
- Search for “Storage Account” in the home and create a storage
- In the storage account let’s create a container and name it as “adfdemo”



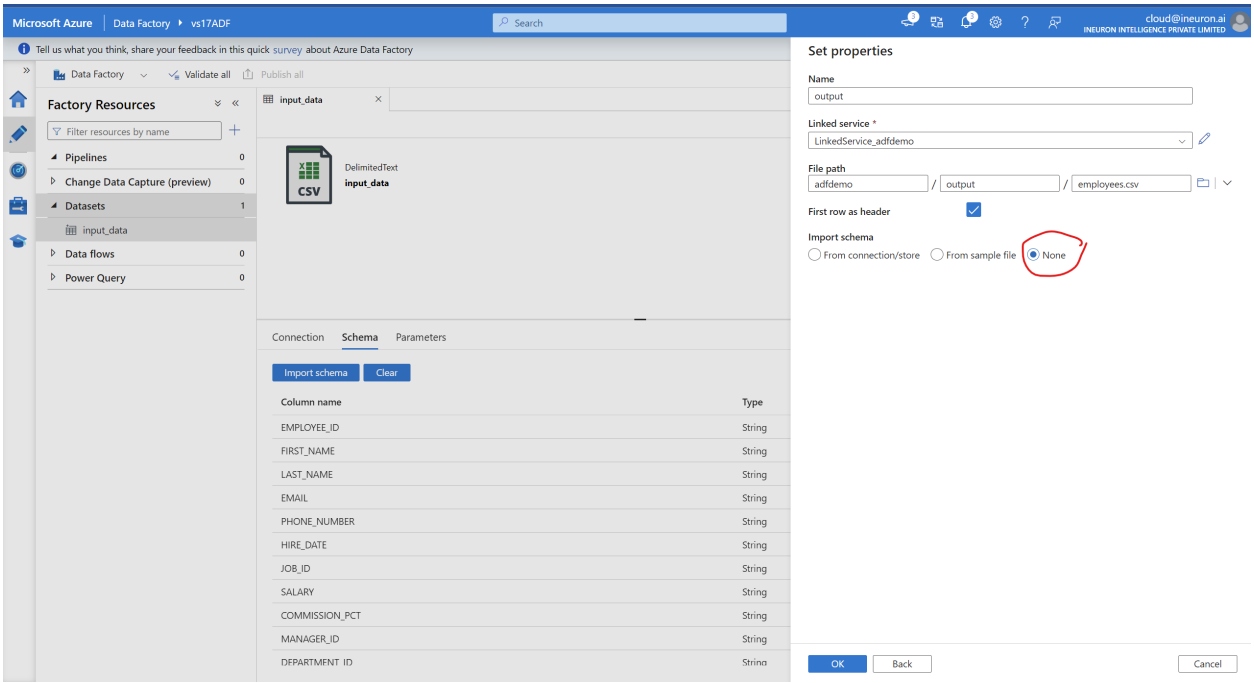
- Now let's place some data inside the container. Create an input folder and place the data inside the input folder.
- Now, let's create ADF and go into "linked services" and create a connection with the "Azure Blob Storage"



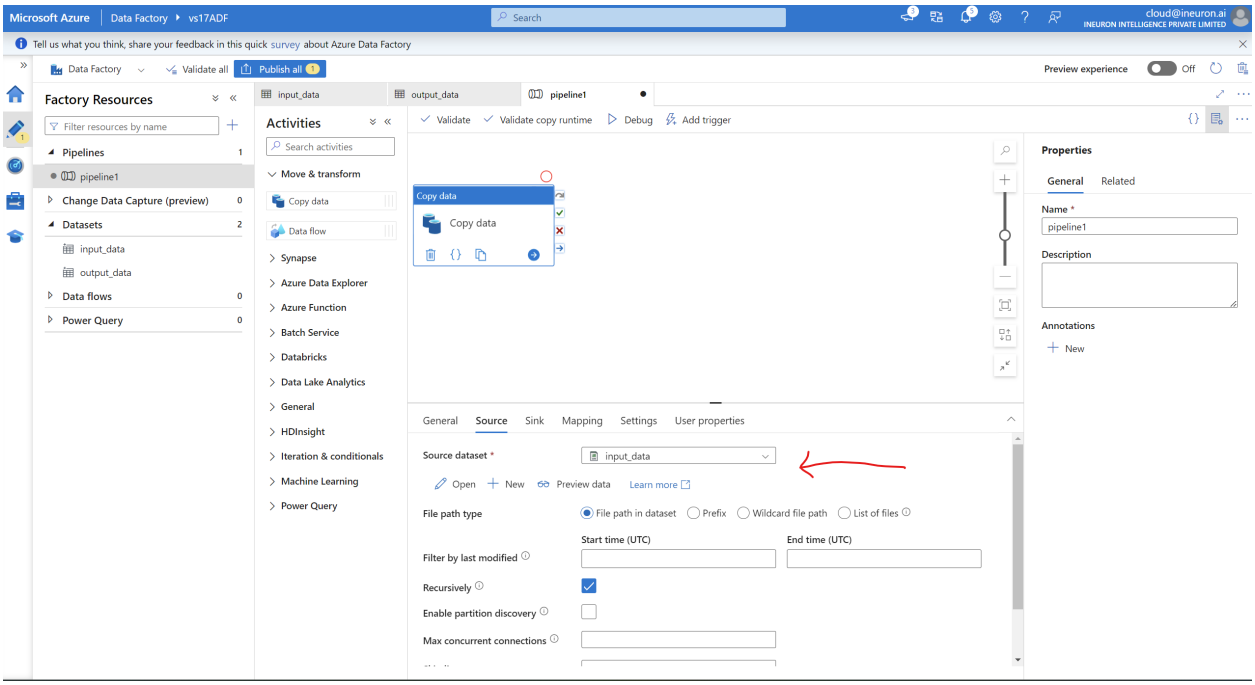
- Create a dataset for the input dataset.

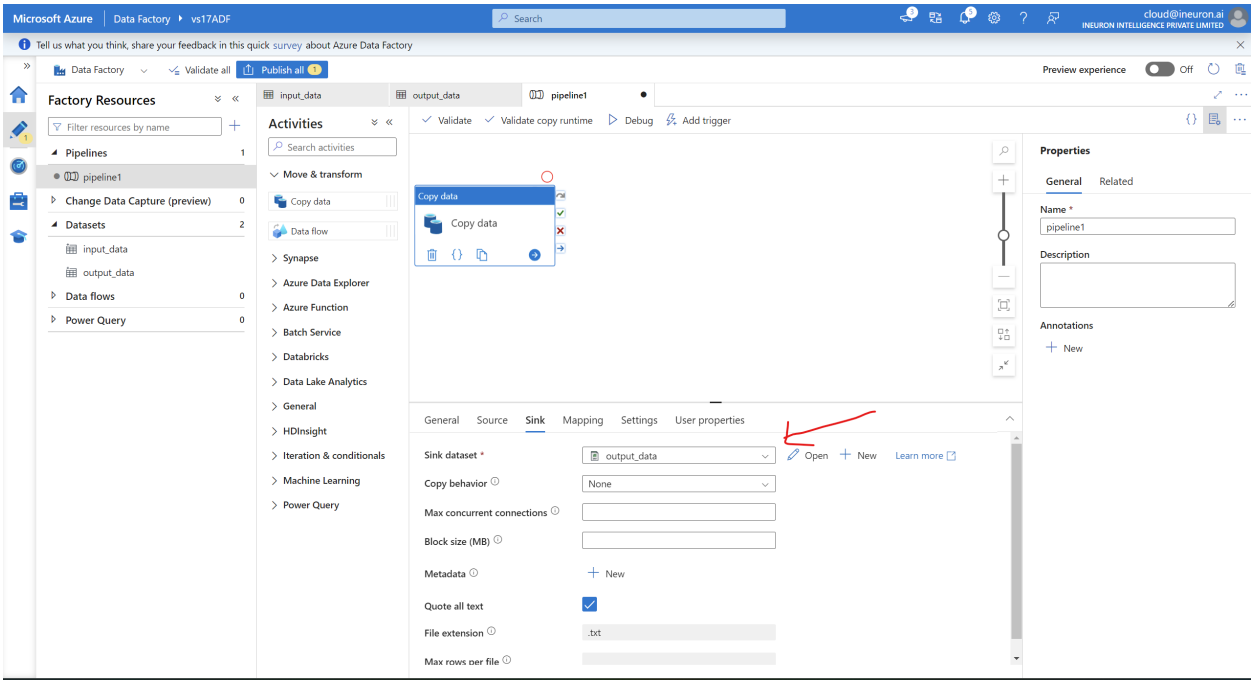


- Create a dataset for the output folder



- Create a pipeline and choose “copy data”





- Click on validate. Once validation is passed click on “Debug” to run the pipeline.
- To run the pipeline we have to trigger the pipeline.

Different ways to create ADF

There are multiple options using which we can create our ADF

- Azure Portal UI
- Azure Powershell → Installation required
- .NET
- Python
- Rest APIs
- Resource Manager Template → Using json config file

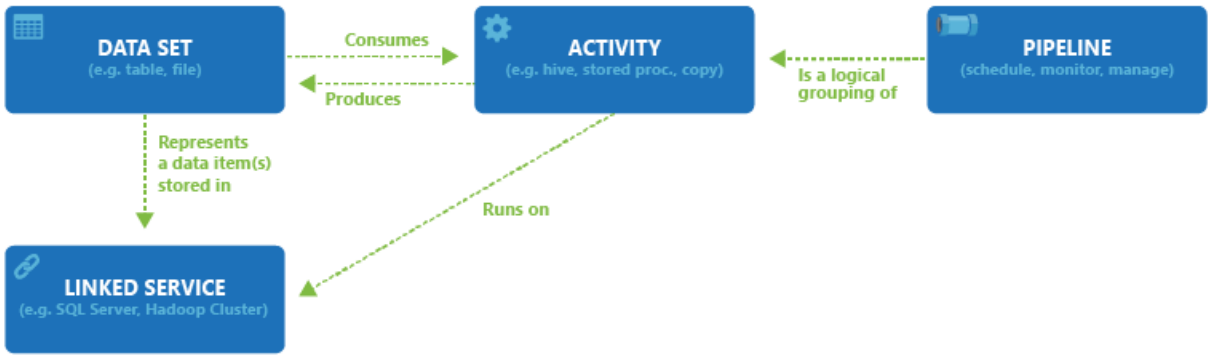
Documentation

Pipelines and activities in ADF

Documentation

Linked Services and Datasets

Linked services are used to connect other resources with ADF. Linked services are like connection strings for resources to connect. Datasets are simply points or references the data, which you want to use in your activity as input or output.



Triggers in ADF

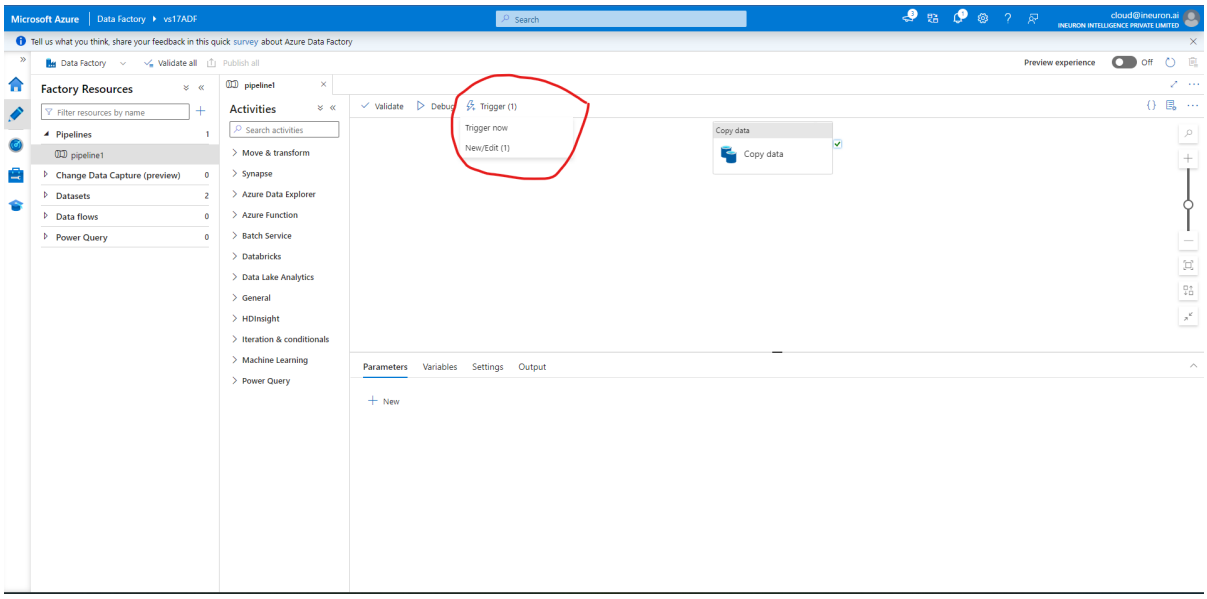
Instead of manually running the ADF pipelines are can make triggers to run our pipeline. Triggers determines when a pipeline execution needs to be kicked off.

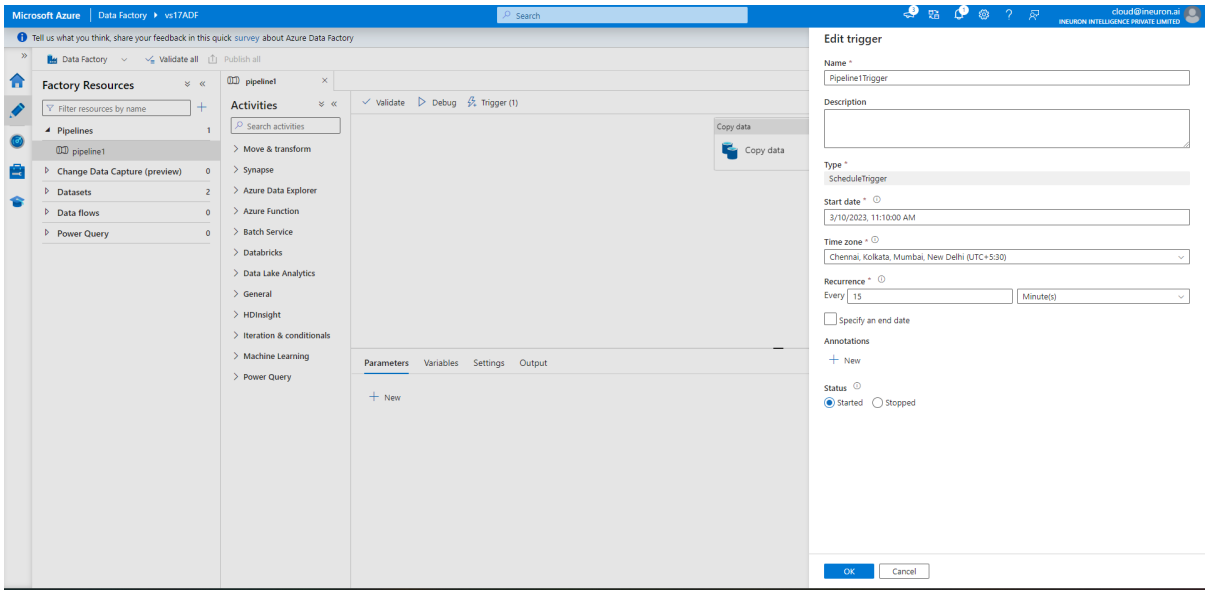
Pipelines and triggers have a many to many relationship, Multiple triggers can kick off a single pipeline or a single trigger can kick off multiple pipelines. Tumbling window trigger is an exception.

There 4 types of triggers

- Schedule Trigger - It invokes a pipeline on a wall-clock schedule
- Tumbling Window Trigger - It operators on a periodic interval, while also retaining state
- Event-based Trigger - It responds to an event
- Custom Trigger - It can be created according to our custom conditions

Creating Schedule Trigger

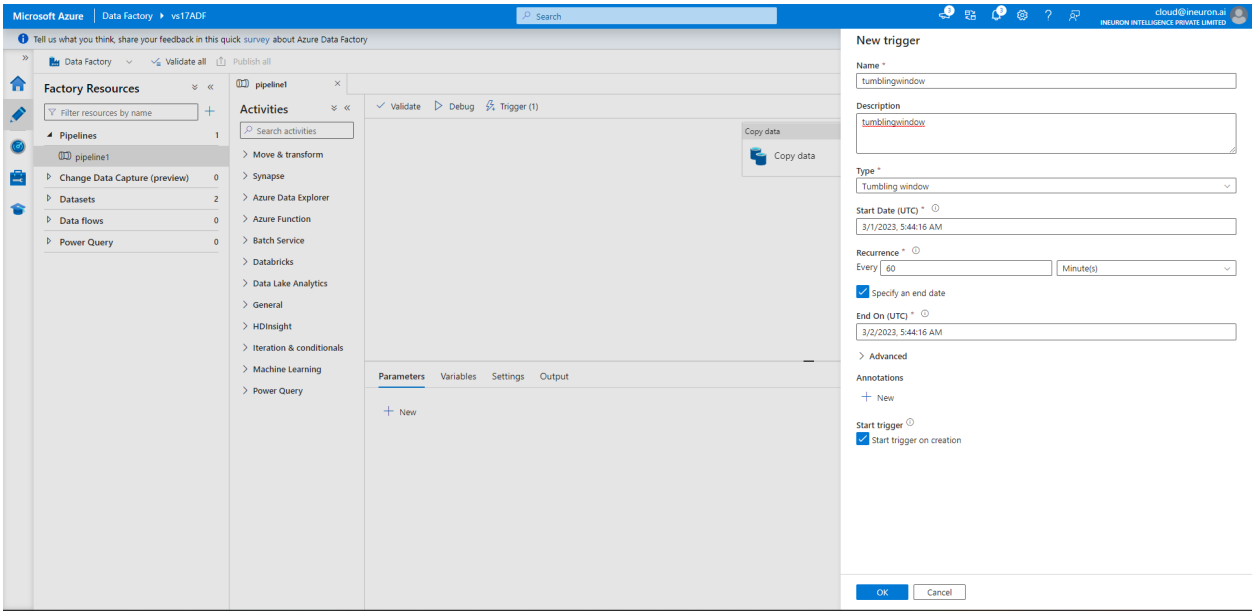




Creating Tumbling Window Trigger

Tumbling window triggers are a type of trigger that fires at a periodic time interval from a specified start time, while retaining the state. A tumbling window trigger has a one-to-one relationship with a pipeline and can only reference a singular pipeline.

Advantage of Tumbling window is that we can even use it for the backfill scenarios (loading historical data). Schedule trigger won't run with the historical data but we can do it with the help of tumbling window.



It will create 24 windows of 60 minutes

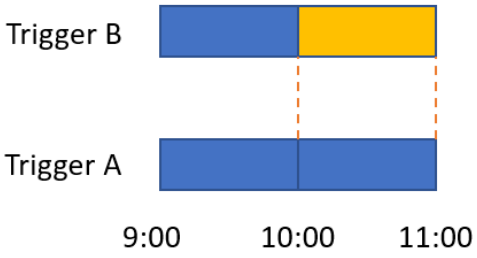
```
3rd march 5:44:16 AM - 3rd march 6:44:16 AM
3rd march 6:44:16 AM - 3rd march 7:44:16 AM
.
.
.
.
4th march 4:44:16 AM - 4th march 5:44:16 AM
```

Let's take scenario where we have two pipelines.

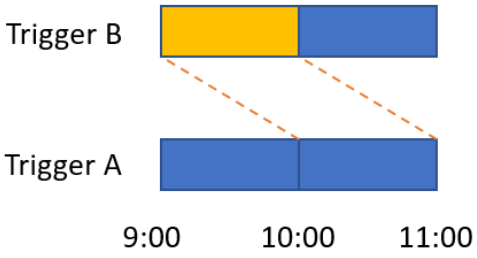
- HourlyLogsPipeline → process logs for every hour and load data into SQL
- HourlyDataProcessPipeline → takes logs hourly data from SQL and combines customer data and then do some transformation.

Here, pipeline 2 is dependent on pipeline 1 so we can make sure that pipeline 2 only runs if pipeline 1 has run successfully.

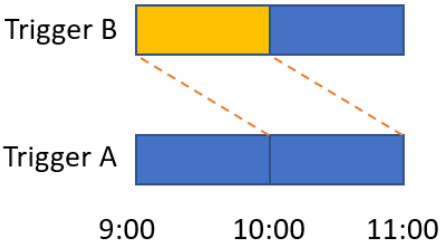
Let's create 2 pipelines. Once pipeline has been created we can create triggers and the make the Hourly Data Process Pipeline depend on Hourly Logs Pipeline.



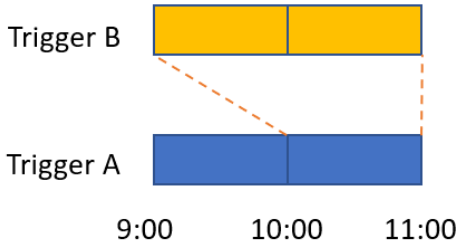
Dependency: A -> B
Offset: 0



Dependency: A -> B
Offset: -1 hour



Dependency: A -> B
Offset: -1 hour
Size: unspecified



Dependency: A -> B
Offset: -1 hour
Size: 2 hours

Documentation

Creating Storage Events Triggers

An event based trigger runs the pipeline in response to an event. Events can be arrival of a file, deletion of file, etc in Azure Blob Storage. Data integration scenarios often require ADF customers to trigger pipeline based on events such as the arrival or deletion of a file in your Azure Storage account.

Documentation

Creating Custom Event Triggers

Documentation

Application

Azure functions

Azure functions is a serverless compute service that lets you run event-triggered code without having to explicitly provision or manage infrastructure. It lets us run small pieces of code called functions without worrying about the application infrastructure.

Features of Azure Functions:

- Serverless application
- Choice of languages - C, C++, C#, Java, JavaScript, Python
- Pay-per-use pricing model
- Bring your own dependencies
- Integrated Security
- Simplified Integration
- Flexible Development

What we can do with Azure functions?

- Azure functions is a great solution for processing bulk data, integrating systems, working with IoT and building simple APIs and micro services.
- We can run Azure functions on various events and triggers.
    1. On HTTP request
    2. On schedule timer

3. On document addition or modification on any other Azure service

Creating Azure function using Visual Studio

Download visual studio and in the Workloads tab select “Azure Development” and “Python Development” while installing.