Azure Data Engineering with Azure Data Factory

Capgemini

Agenda

- What is Cloud Computing?
- Subscription and Resource Group
- Azure Data Factory (v2)

What is Cloud Computing?

Cloud computing is the on-demand availability of computer system resources such as servers, storage, databases, networking, software, analytics, and intelligence—over the Internet ("the cloud") to offer faster innovation, flexible resources, and economies of scale

Typical services provided by the cloud service provider:

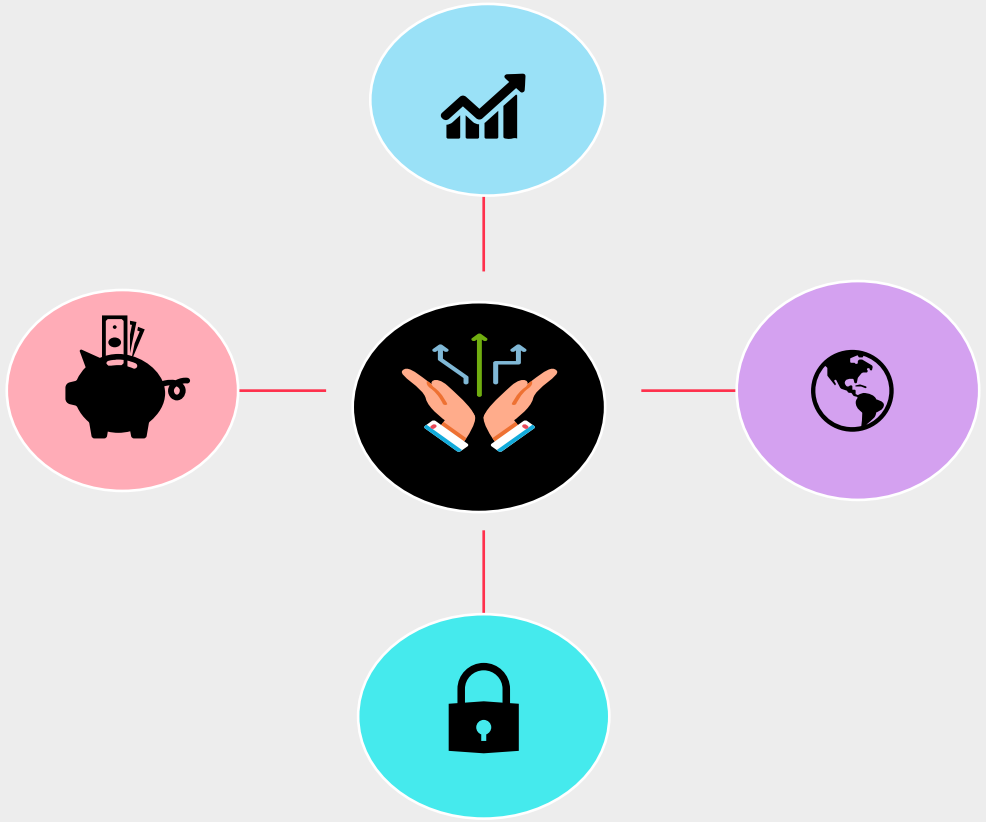
- **Compute -** virtual machines and servers
- **Storage –** Databases and file systems
- **Networking –** Secure connection between cloud & on-premise domains
- **Analytics -** Visualization of telemetry and performance data

Source: https://www.cloudns.net/blog/microsoft-azure-iaas-paas-saas/

Why Cloud Computing?

Scalable

Seamless scale in computation both horizontally and vertically



Cost Effective

'Pay As you Go' or 'Consumption' based billing

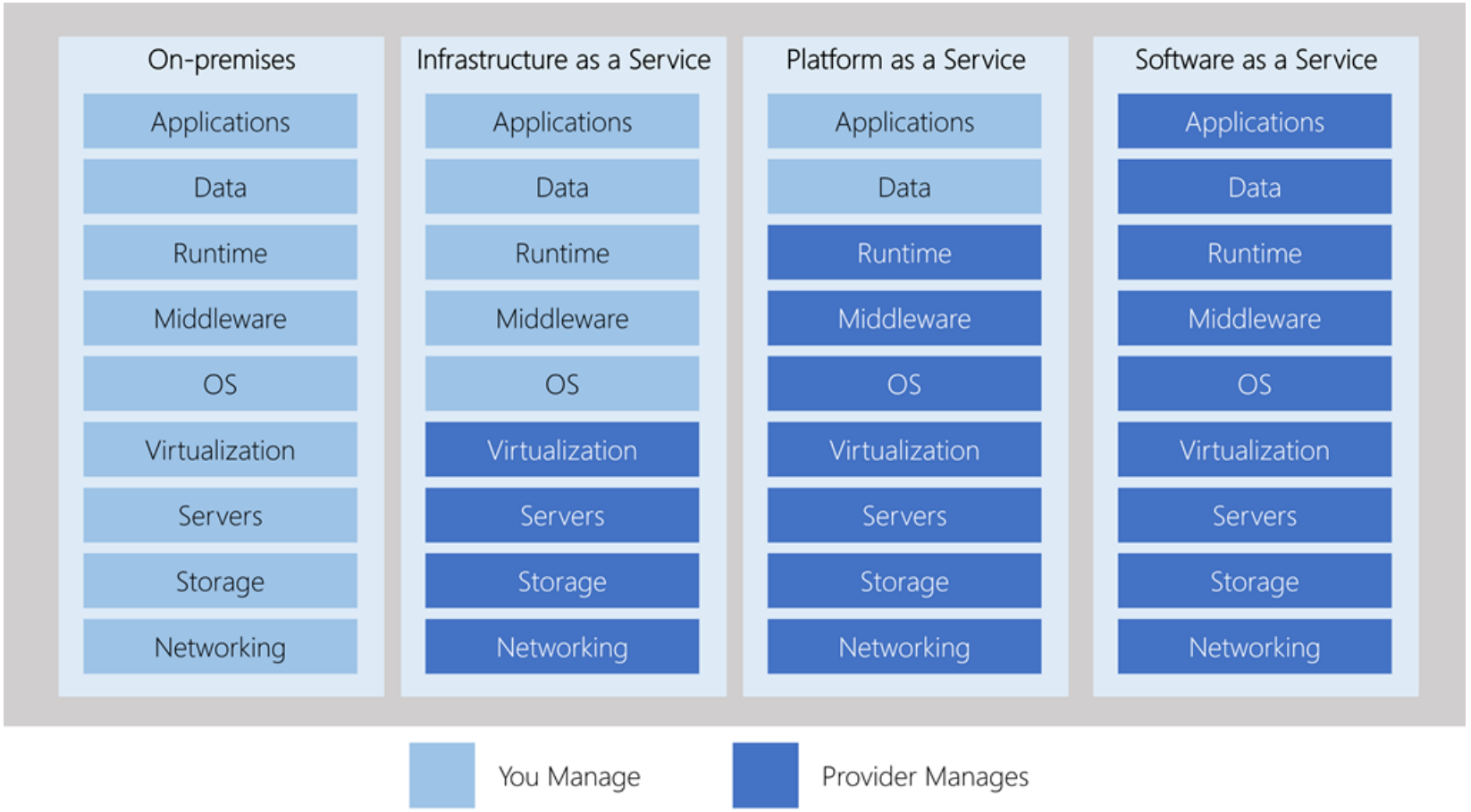
Global Distribution

Interconnected, low latency datacenters worldwide that ensures the best response and end user performance

Secured

Multi layered security models that ensures applications and data is compliant and prevents breaches

Categories of Services on Cloud Platforms



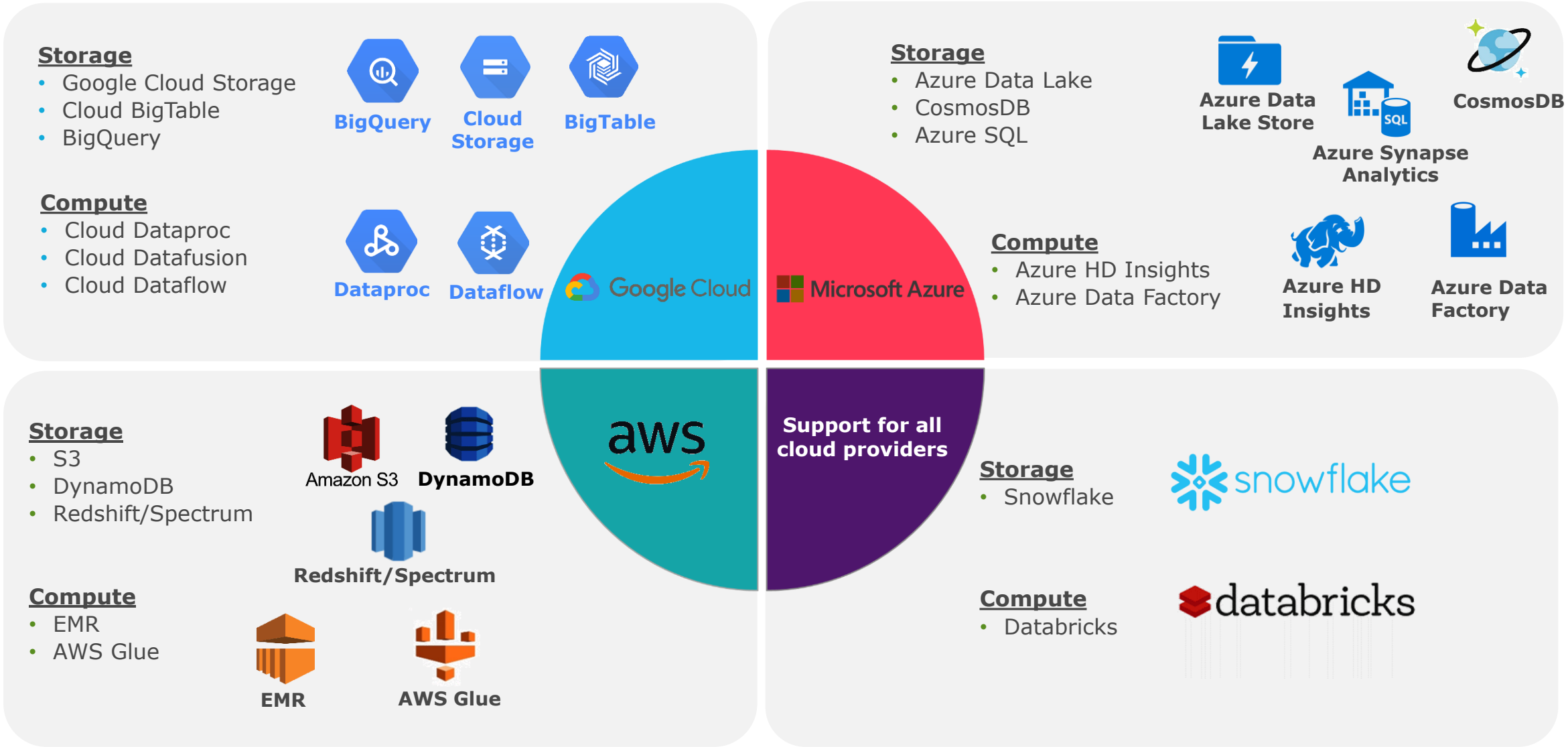
Infrastructure as a service (IaaS)
Infrastructure as a Service is the most flexible category of cloud services. It aims to give you the most control over the provided hardware that runs your application (IT infrastructure servers and virtual machines (VMs), storage, and operating systems). Instead of buying hardware, with IaaS, you rent it. It's an instant computing infrastructure, provisioned and managed over the internet.

Platform as a service (PaaS)
PaaS provides an environment for building, testing, and deploying software applications. The goal of PaaS is to help you create an application quickly without managing the underlying infrastructure. For example, when deploying a web application using PaaS, you don't have to install an operating system, web server, or even system updates.

Software as a service (SaaS)
SaaS is software that is centrally hosted and managed for the end customer. It is usually based on an architecture where one version of the application is used for all customers, and licensed through a monthly or annual subscription. Office 365, Skype, and Dynamics CRM Online are perfect examples of SaaS software.

Source:
https://docs.microsoft.com/en-us/learn/modules/principles-cloud-computing/5-types-of-cloud-services

Leading Cloud Service Providers

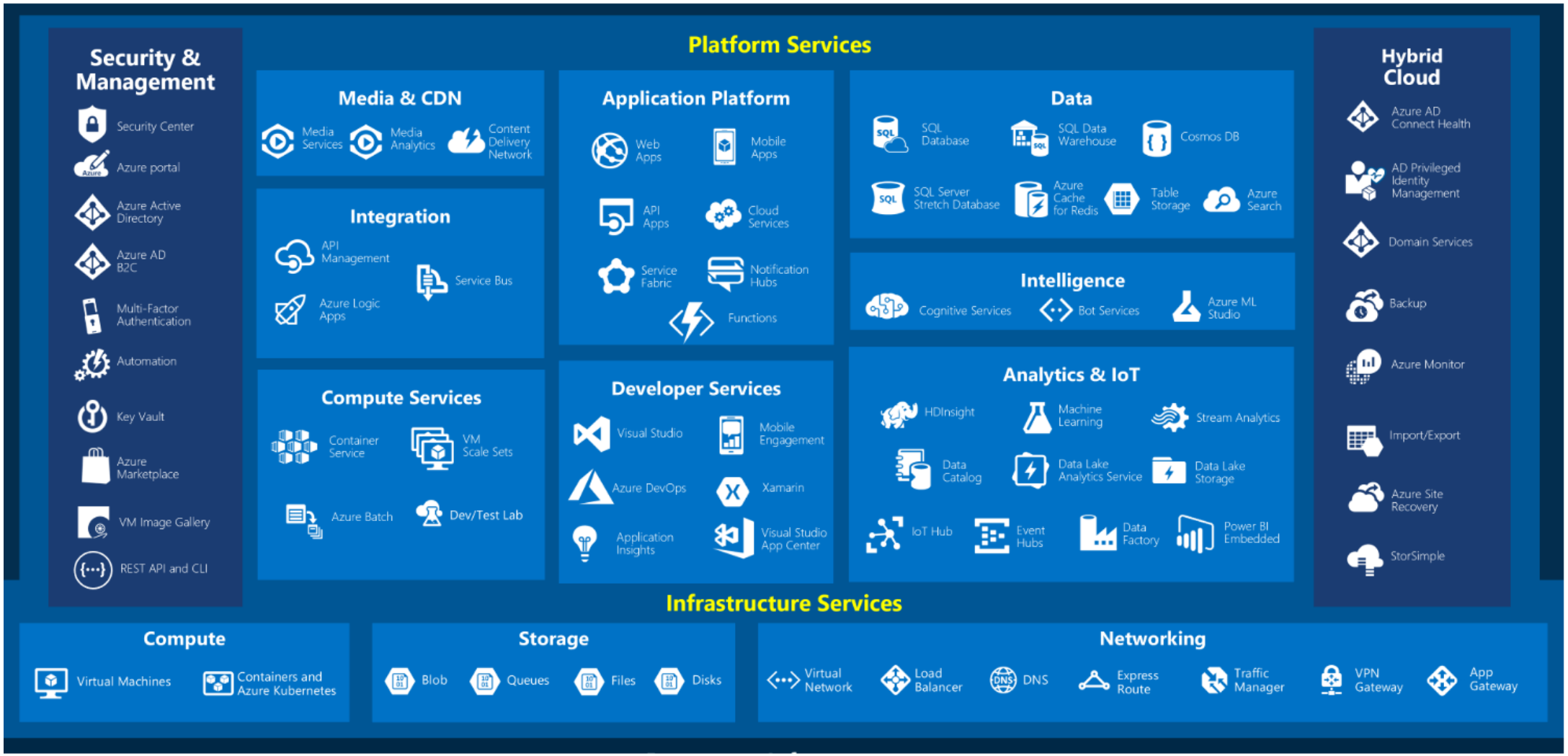


Microsoft Azure
Azure is Microsoft's cloud computing platform. Azure provides over 100 services that enable you to do everything from running your existing applications on virtual machines to exploring new software paradigms such as intelligent bots and mixed reality.

Here are just a few kinds of services you'll find on Azure:

- **Compute** services such as VMs and containers that can run your applications
- **Database** services that provide both relational and NoSQL choices
- **Identity** services that help you authenticate and protect your users
- **Networking** services that connect your datacenter to the cloud, provide high availability or host your DNS domain
- **Storage** solutions that can accommodate massive amounts of both structured and unstructured data
- **AI and machine-learning** services can analyze data, text, images, comprehend speech, and make predictions using data — changing the world of agriculture, healthcare, and much more.

Tour of Services on Azure

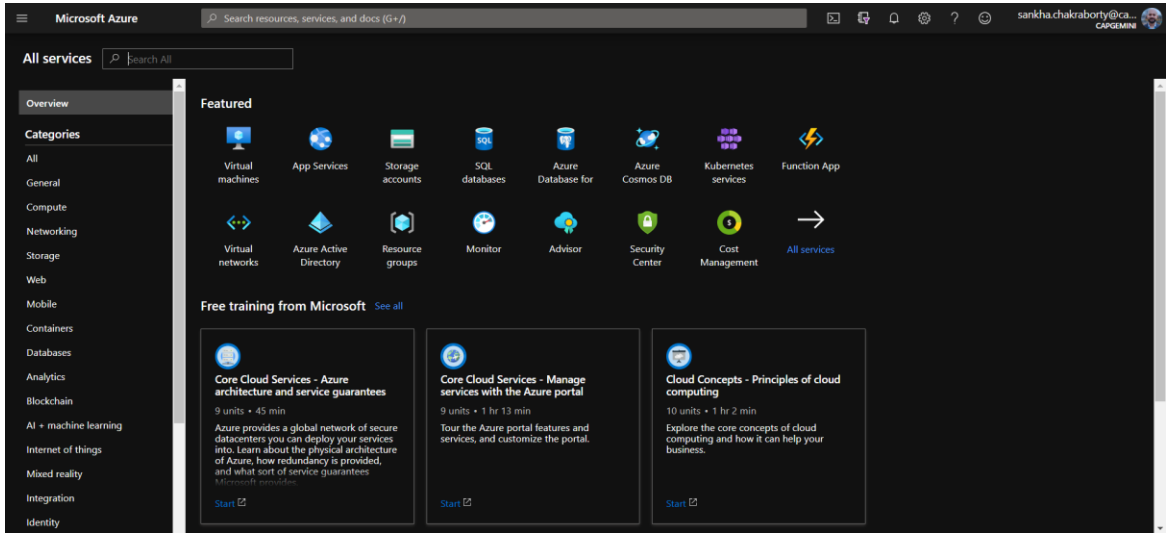


Source:
https://docs.microsoft.com/en-us/learn/modules/welcome-to-azure/3-tour-of-azure-services

Azure Management

Portal:

- A graphical user interface (GUI) that facilitates easy interaction with Azure. User can login and start provisioning services at their will



Azure Cloud Shell:

- Azure Cloud Shell is an interactive, authenticated, browser-accessible shell for managing Azure resources. It provides the flexibility of choosing the shell experience that best suits the way you work, either Bash or PowerShell.

Azure PowerShell:

- Cross-platform version of Windows PowerShell that can run Wiindows, macOS or Linux

```
New-AzVM `
    -ResourceGroupName "MyResourceGroup" `
    -Name "TestVm" `
    -Image "UbuntuLTS" `
```

Azure CLI:

- Azure CLI is a cross-platform command-line program that connects to Azure and executes administrative commands on Azure resources

```
az vm create \
  --resource-group MyResourceGroup \
  --name TestVm \
  --image UbuntuLTS \
  --generate-ssh-keys \
```

Resource Management on Azure

Subscription

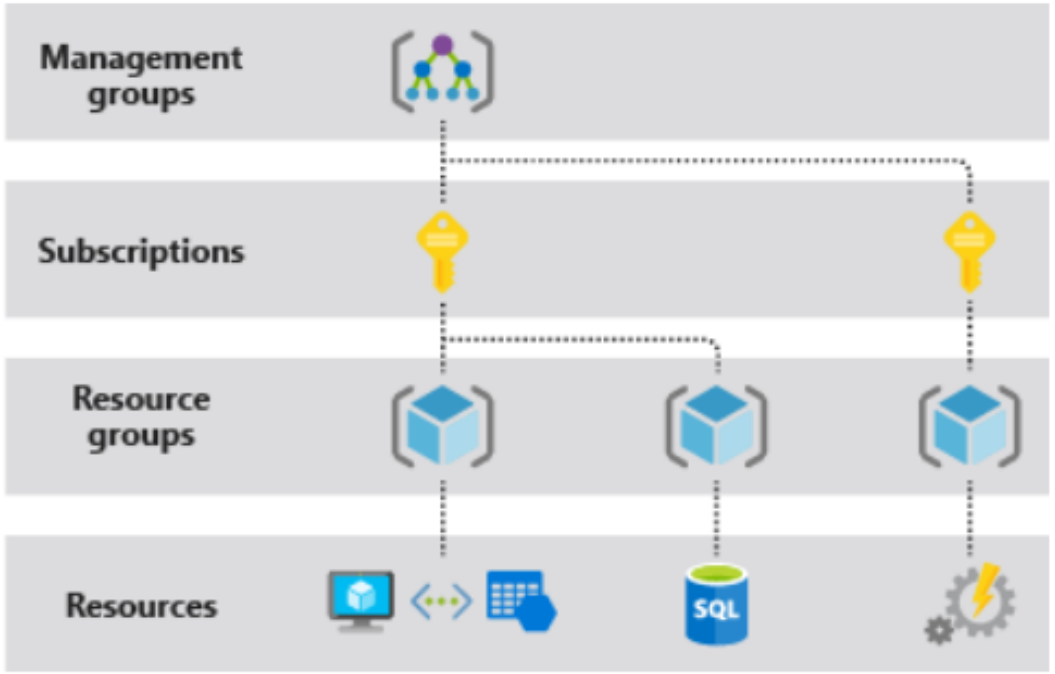
Azure is a great reservoir of resources that your organization can use to deploy applications upon and the cloud is focused around pooling resources together. However, organizations need to be able to split resources up based on cost centres. The development team will be using resources for building new apps, as well as maybe an e-commerce team for production uses. Subscriptions allow for a single Azure instance to separate these costs, and bill to different teams

Resource Group

A container that holds related resources for an Azure solution. The resource group includes those resources that you want to manage as a group. You decide which resources belong in a resource group based on what makes the most sense for your organization

Management Group

A governance framework for managing efficiently manage access, policies, and compliance for those subscriptions. organize subscriptions into containers called "management groups" and apply your governance conditions to the management groups



Source:
https://docs.microsoft.com/en-us/azure/azure-resource-manager/management/overview#resource-groups

Subscription & Resource Group

Subscription

Azure is a great reservoir of resources that your organization can use to deploy applications upon and the cloud is focused around pooling resources together. However, organizations need to be able to split resources up based on cost centres. The development team will be using resources for building new apps, as well as maybe an e-commerce team for production uses. Subscriptions allow for a single Azure instance to separate these costs, and bill to different teams.

Resource Group

Azure Resource Manager is the deployment and management service for Azure. It provides a management layer that enables you to create, update, and delete resources in your Azure subscription. You use management features, like access control, locks, and tags, to secure and organize your resources after deployment.

Example of Azure Compute Services

Azure compute is an on-demand computing service for running cloud-based applications. It provides computing resources like multi-core processors and supercomputers via virtual machines and containers. It also provides serverless computing to run apps without requiring infrastructure setup or configuration.

Virtual Machines

Provision Linux and Windows virtual machines in seconds with the configurations of your choice

Serverless

Accelerate app development using an event-driven, serverless architecture

Container Instances

Containerized apps and easily run containers with a single command

App Service

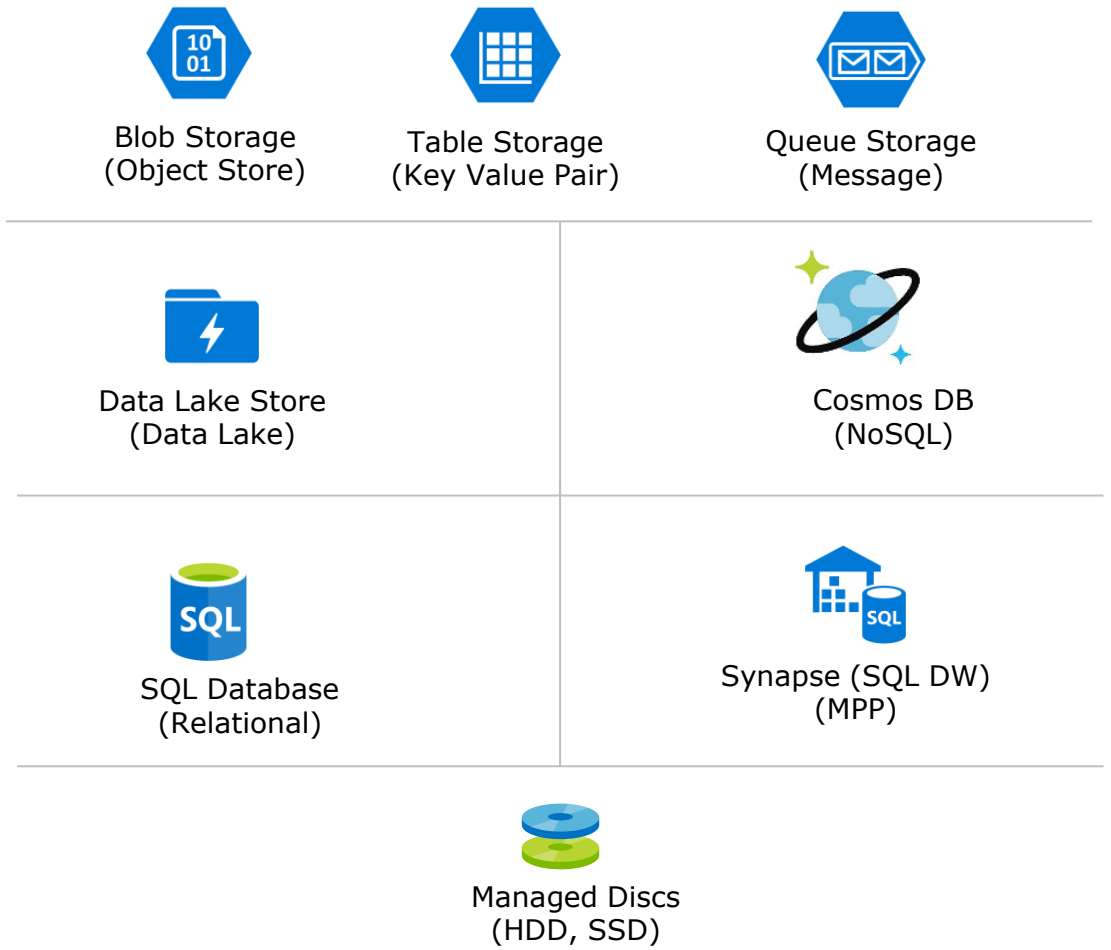
Quickly create cloud apps for web and mobile with fully managed platform

Example of Azure Storage Services

The Azure data storage options are cloud-based, secure, and scalable. Its features address the key challenges of cloud storage and provide you with a reliable and durable storage solution

Here are some of the important benefits of Azure data storage:

- Automated backup and recovery
- Replication across the globe
- Support for data analytics
- Multiple data types
- Data storage in virtual disks
- Storage tiers



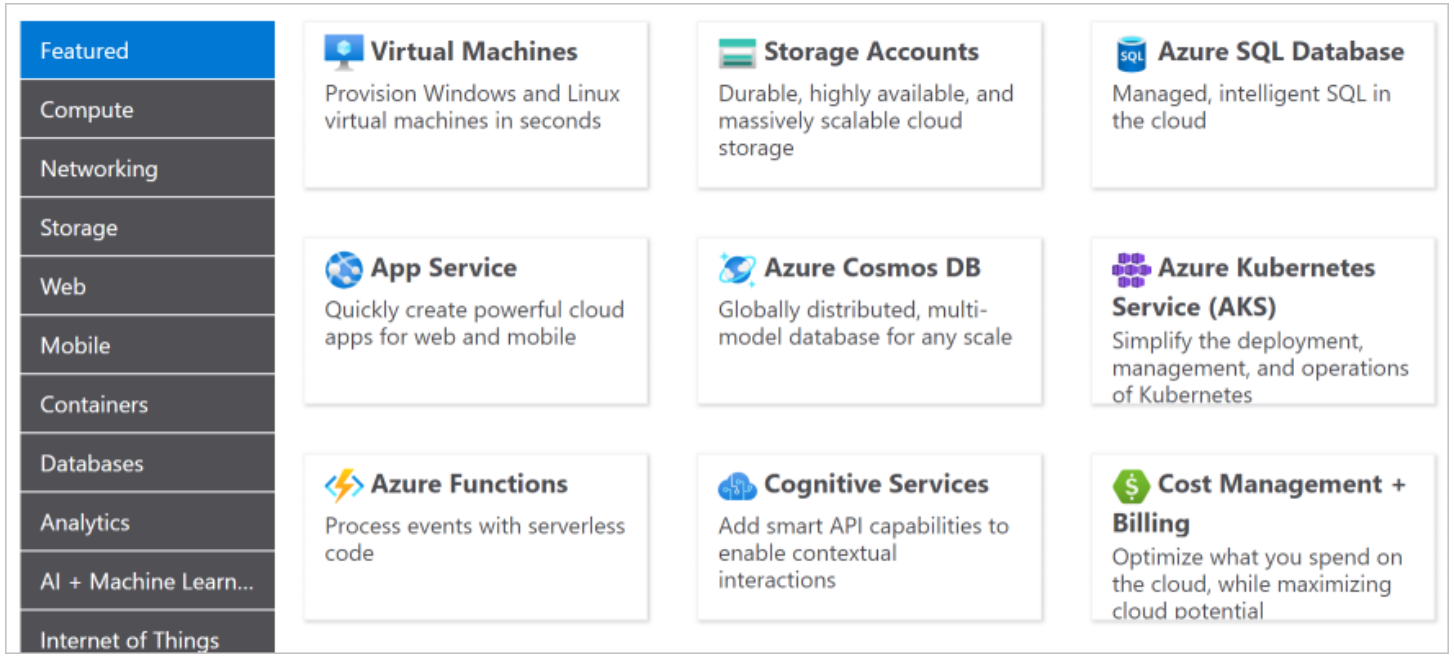
Cost and Spend Optimization on Azure

Usage meters:

When you provision an Azure resource, Azure creates one or more-meter instances for that resource. The meters track the resources' usage and generate a usage record that is used to calculate your bill.

- For example, a single virtual machine that you provision in Azure might have the following meters tracking its usage:
  - Compute Hours
  - IP Address Hours
  - Data Transfer In
  - Data Transfer Out
  - Standard Managed Disk
  - Standard Managed Disk Operations
  - Standard IO-Disk
  - Standard IO-Block Blob Read
  - Standard IO-Block Blob Write
  - Standard IO-Block Blob Delete

Pricing Calculator



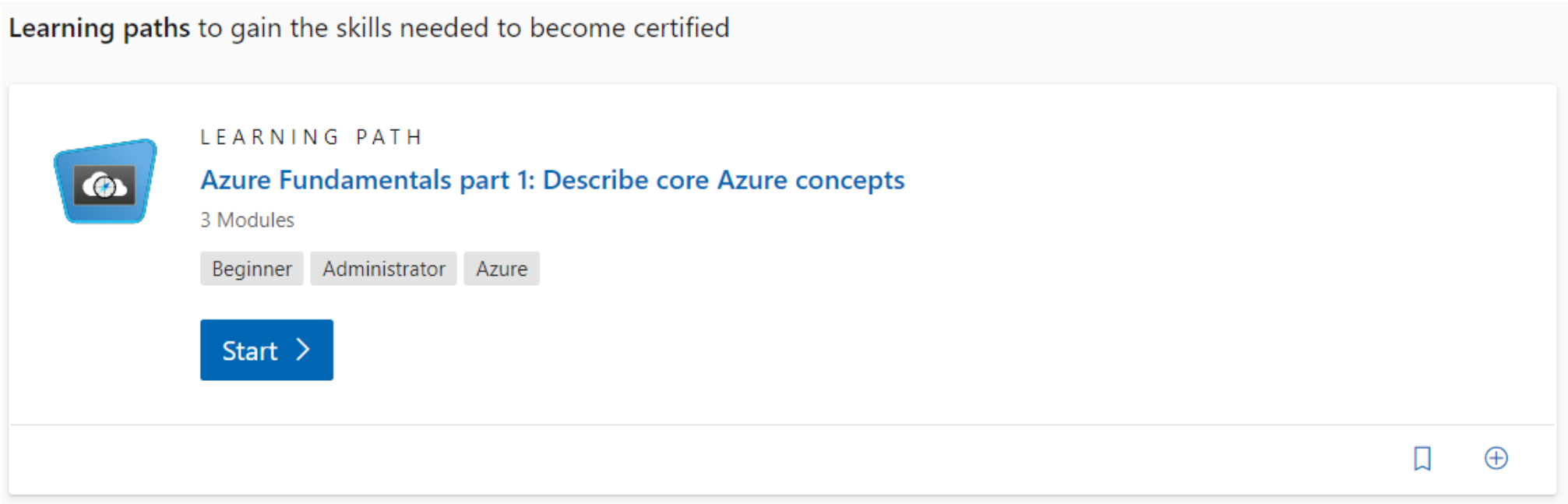
Source:
https://azure.microsoft.com/en-us/pricing/calculator/

P.S: Always stop/terminate/pause your compute instances after you are done working on it to save cost

Certification – AZ-900

- To clear all the functionals of Azure Cloud Computing it is highly recommended that you do the AZ-900 certification.
- There is free course available from Microsoft for the exam. It is 6 sections. Please follow the link below for more details.

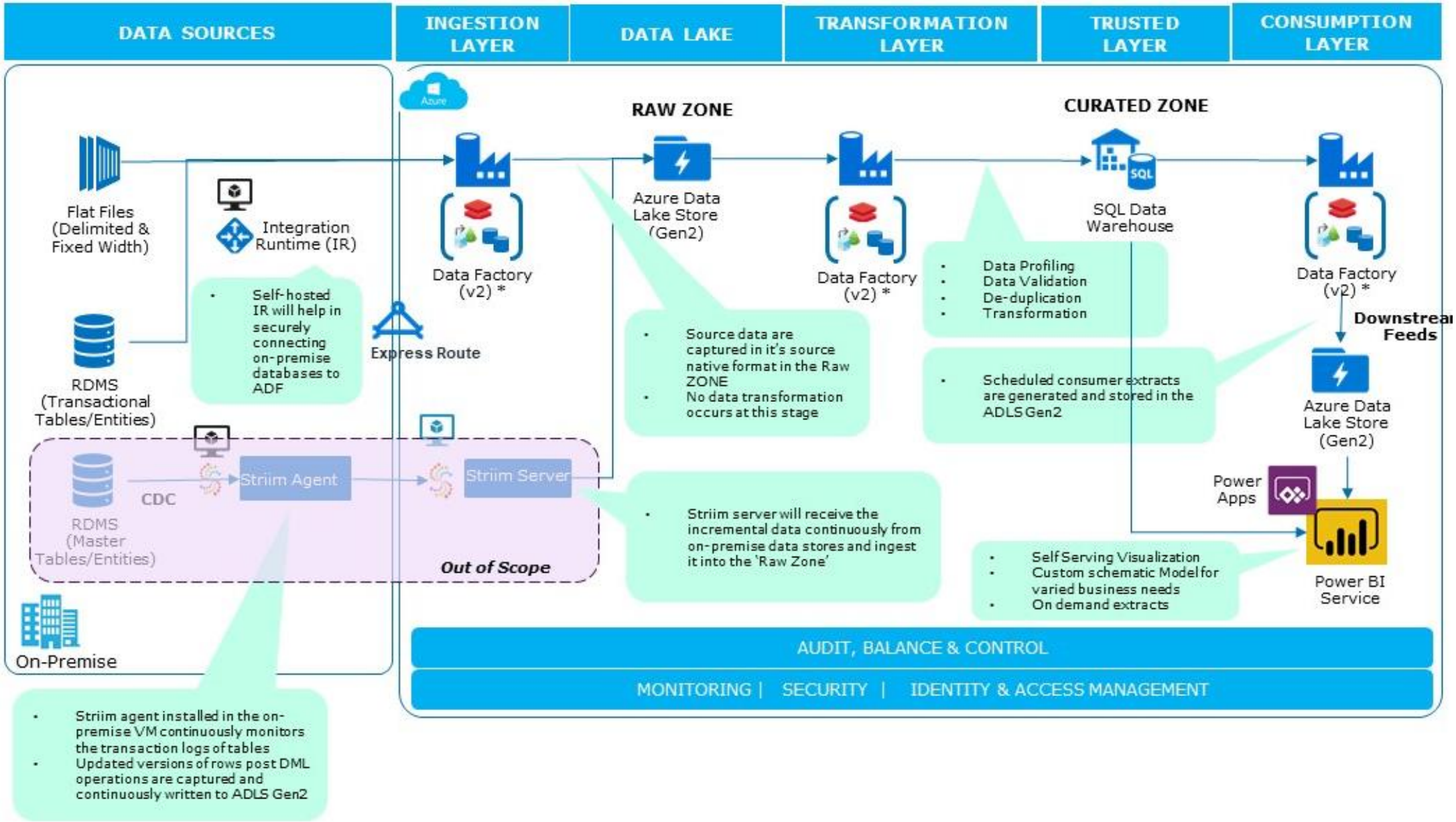
Exam AZ-900: Microsoft Azure Fundamentals



Reference Architecture

Capgemini

Architecture derived for client



LAB

- Creating Office 365 subscription
- Walkthrough of Azure Portal

Azure Storage, Azure Data Lake Service

Capgemini

Azure Data Lake Storage – Gen2

- Azure Data Lake Storage Gen2 was released in February 2019

- ADLS Gen2 is built on Azure Blob Storage and combines the object storage and file system paradigms

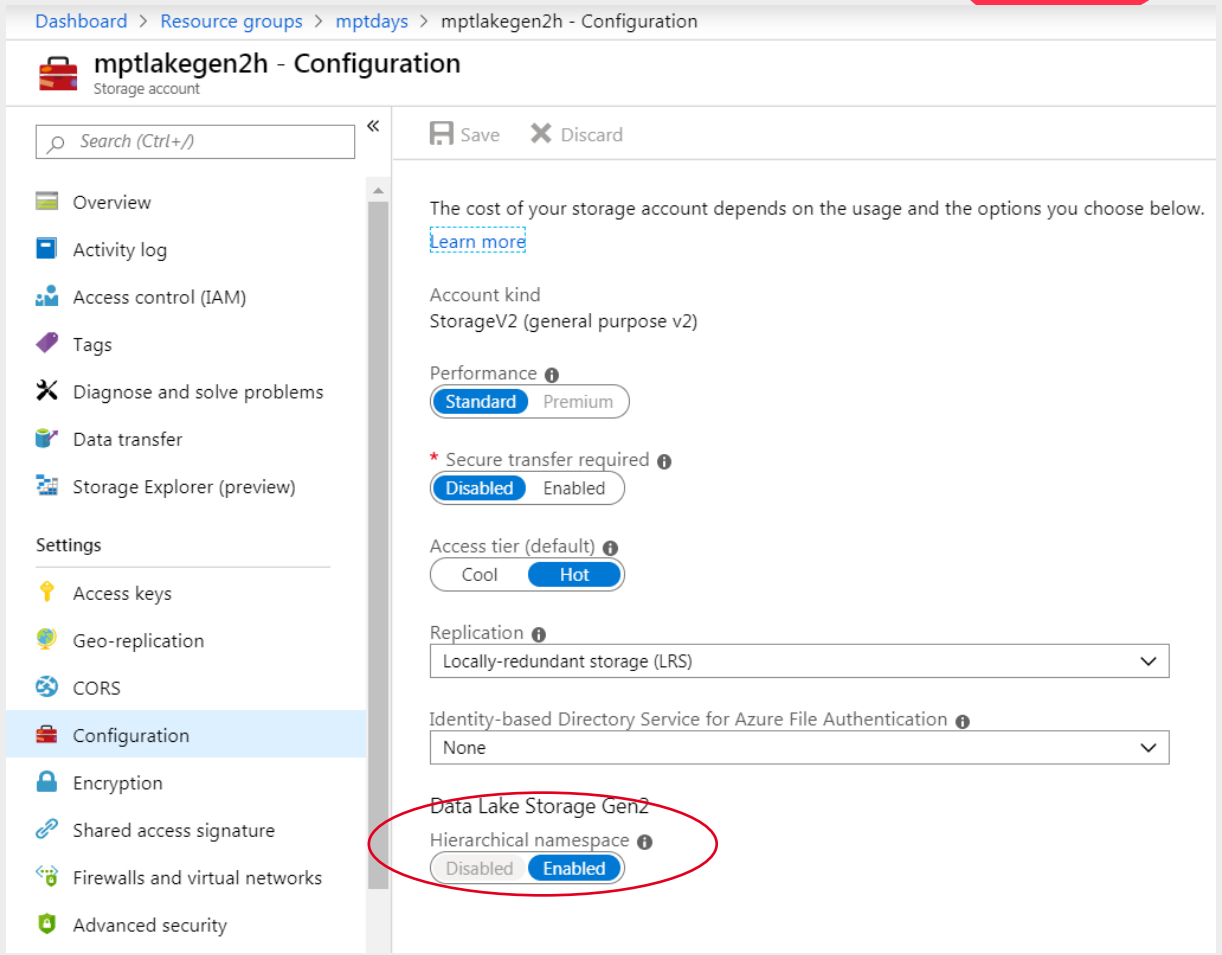
- Best of both worlds
  - From the Blob Storage
    - Cheaper storage (cold and hot storage)
    - Flat namespace storage
  - From Azure Data Lake
    - HDFS (parallel reads and writes)
    - AAD security
- Key Benefits:
  - Performance: hierarchical namespace improves performance of directory management operations
  - Cost effectiveness: ADLS Gen2 is built on top of low-cost Azure Blob Storage
  - Security: Create POSIX permissions on directories or individual files

- Use HDFS or Flat Namespace by connecting to the Storage by using separate drivers:
  - Azure Blob File System driver
  - Hadoop filesystem driver

Creating Azure Data Lake Service (Gen-2)

LAB

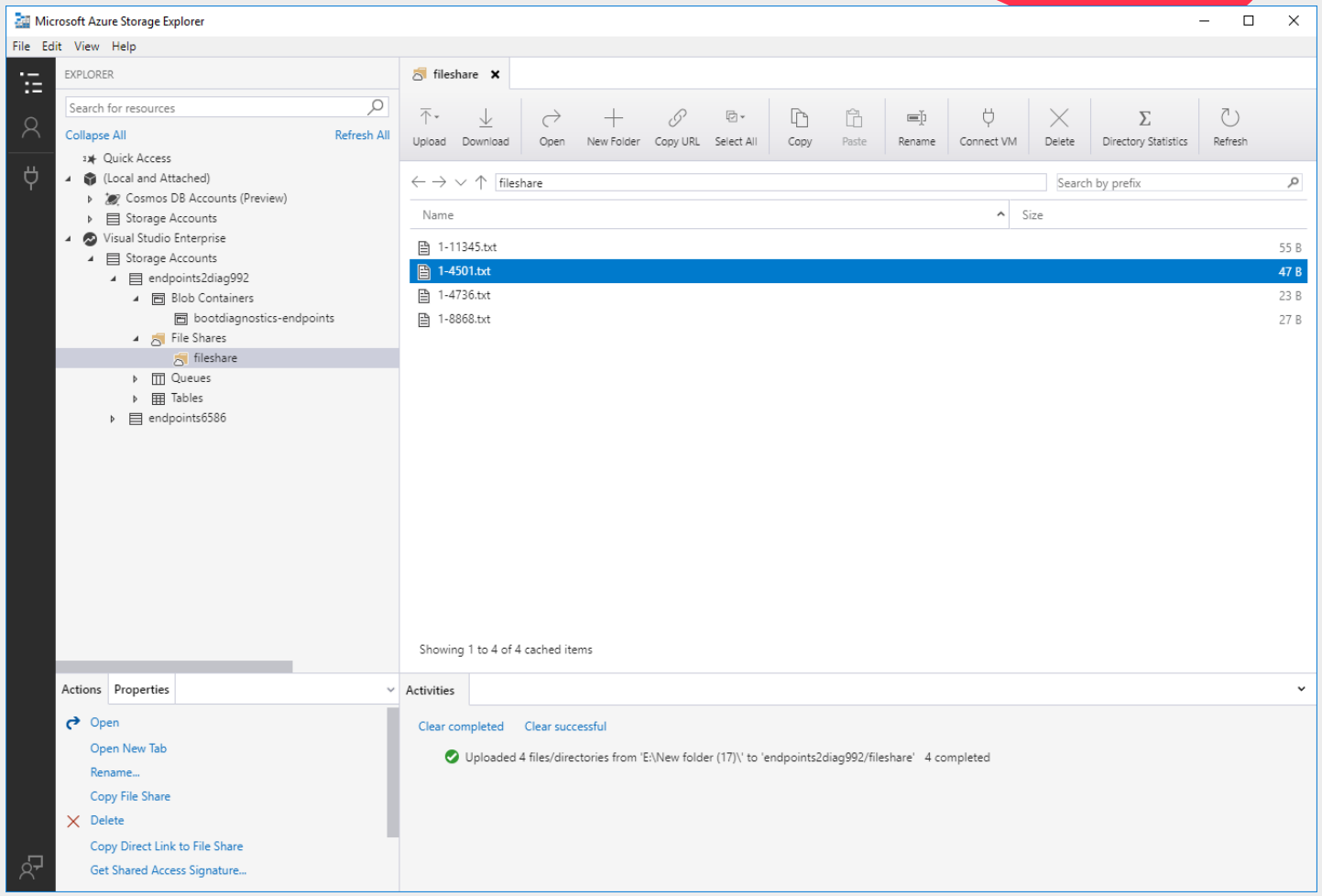
- **Blob Storage**
- ~~**Azure Data Lake Gen1**~~
- **Azure Data Lake Gen2**
  - Azure Data Lake Gen2 is a Blob Storage with Hierarchical Namespace enabled.



Creating Azure Data Lake Service (Gen-2)

LAB

- Downloadable extra tool
  - Available for Windows, Mac, and Linux
- Features are
  - View and edit Blob, Queue, Table, File, Cosmos DB storage and Data Lake Storage
  - Create, delete, view, and edit storage resources
  - Obtain shared access signature (SAS) keys
    - Discussed in session "Design for Security"
  - Manage Snapshots



Picture from docs.microsoft.com

Azure Data Factory V2

Capgemini

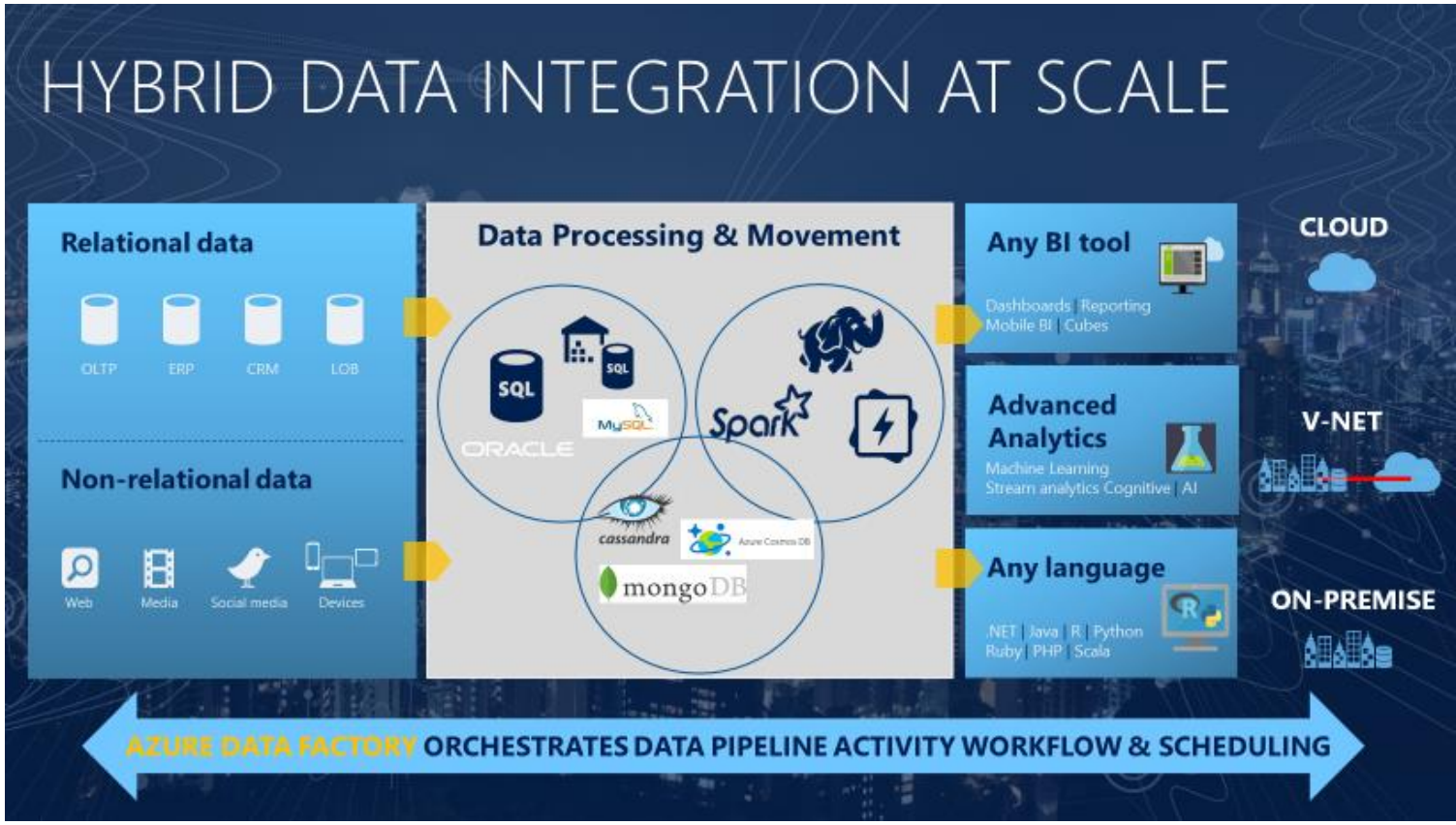
Introduction to Azure Data Factory (v2) [aka. ADFv2]

Azure Data Factory (v2)

Data Factory is the Azure cloud ETL tool which orchestrates data pipeline activities and handles scheduling.

A typical scenario covers

- Connect & collect data
- Transform & enrich data
- Publish data
- Monitor



ADFv2 : Key Concepts (1/6)

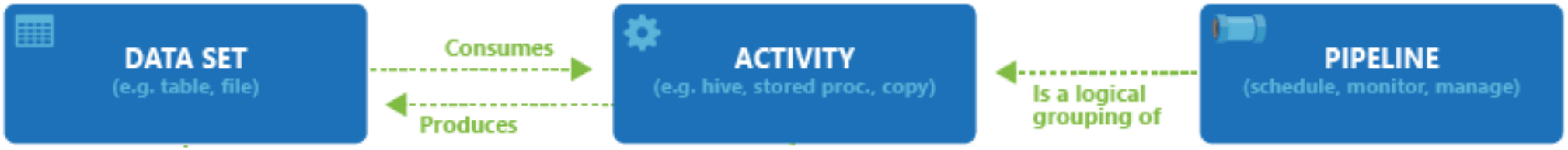
Pipelines and activities in Azure Data Factory

Pipeline

- A pipeline is a logical grouping of activities that together perform a task.
- For example, a pipeline could contain a set of activities that ingest and clean log data, and then kick off a mapping data flow to analyse the log data.
- The pipeline allows you to manage the activities as a set instead of each one individually

Activities

- The activities in a pipeline define actions to perform on your data.
- For example, you may use a copy activity to copy data from an on-premises SQL Server to an Azure Blob Storage. Then, use a data flow activity or a Databricks Notebook activity to process and transform data from the blob storage to an Azure Synapse Analytics pool on top of which business intelligence reporting solutions are built

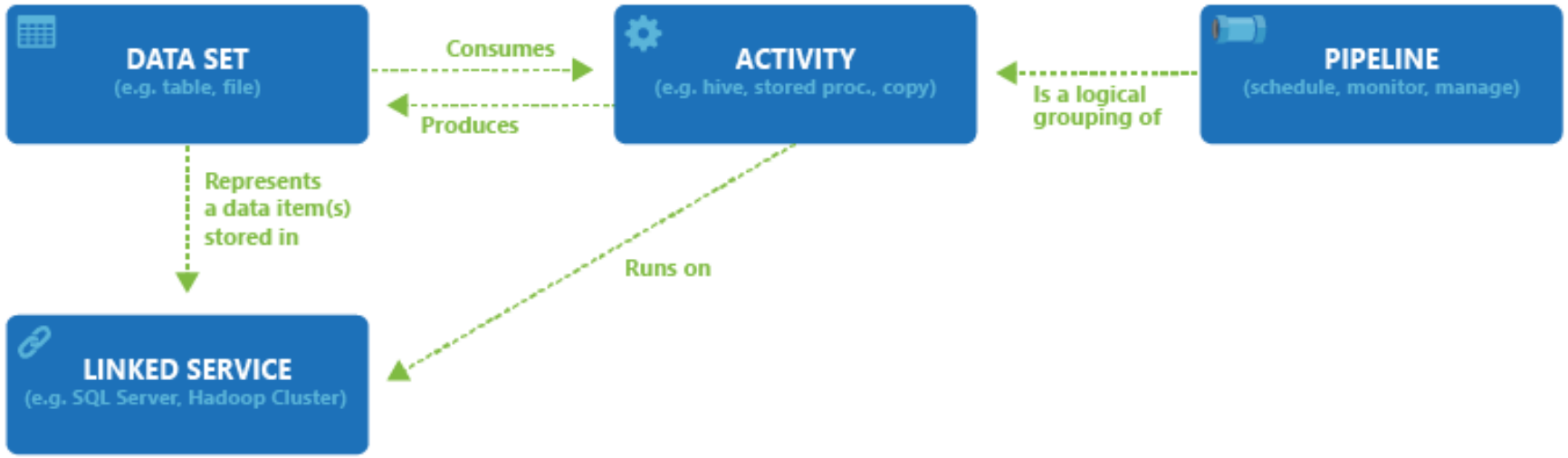


Azure Docs Link: Concepts- Pipelines Activities

ADFv2 : Key Concepts (2/6)

Linked Services

- Linked services are much like connection strings, which define the connection information needed for Data Factory to connect to external resources

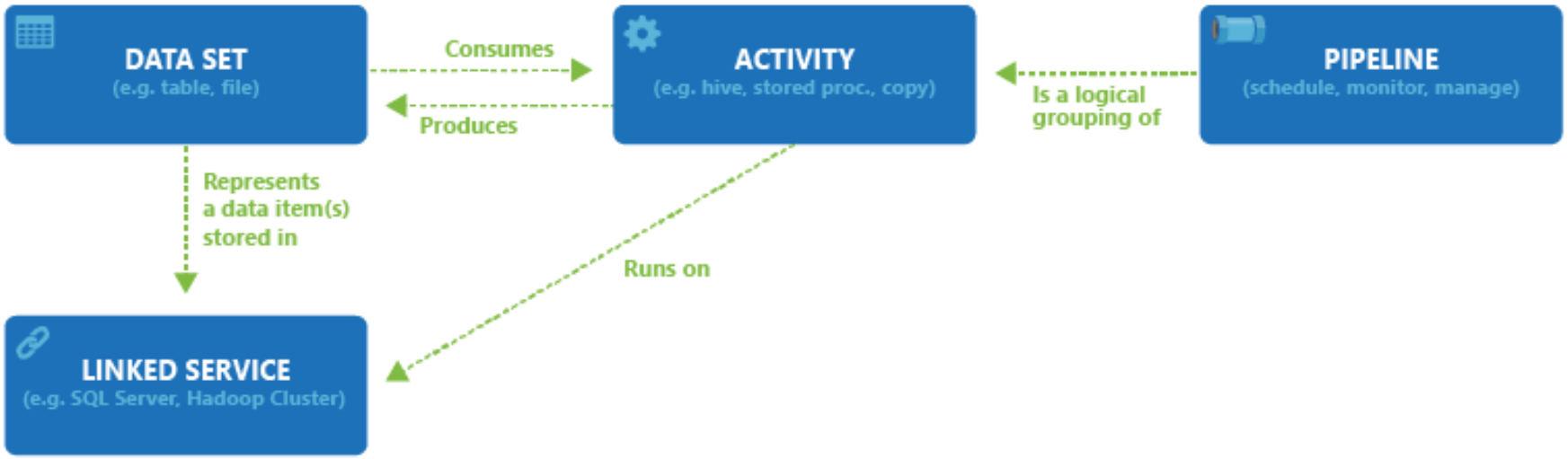


Azure Docs Link: [Concepts- Linked Services]

ADFv2 : Key Concepts (3/6)

Datasets

- Dataset is a named view of data that simply points or references the data you want to use in your activities as inputs and outputs.
- Datasets identify data within different data stores, such as tables, files, folders, and documents. For example, an Azure Blob dataset specifies the blob container and folder in Blob storage from which the activity should read the data.

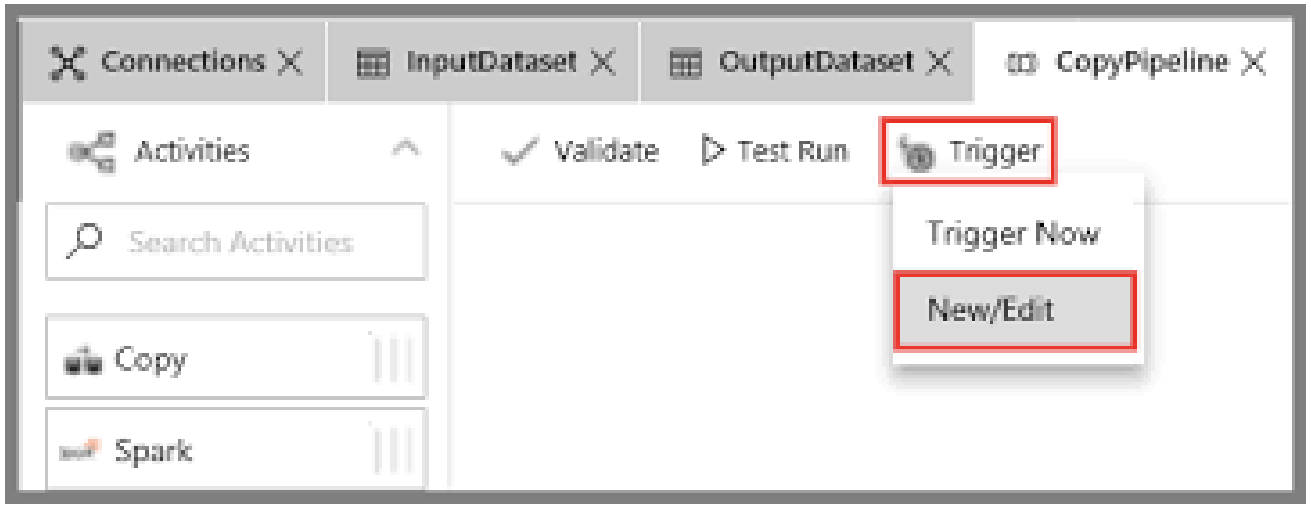


Azure Docs Link: [Concepts- Datasets]

ADFv2 : Key Concepts (4/6)

Pipeline execution and triggers

- A pipeline run in Azure Data Factory defines an instance of a pipeline execution
- Pipeline runs are typically instantiated by passing arguments to parameters that you define in the pipeline
- You can execute a pipeline either manually or by using a trigger. This article provides details about both ways of executing a pipeline.

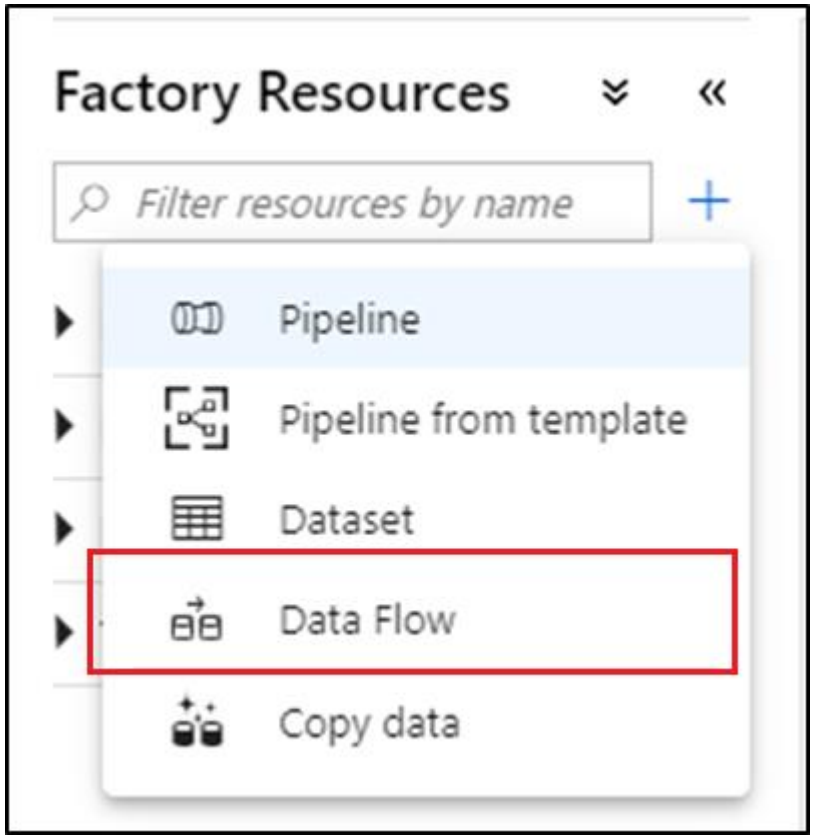


Azure Docs Link: [Concepts- Pipeline Execution Triggers]

ADFv2 : Key Concepts (5/6)

Dataflow

- Mapping data flows are visually designed data transformations in Azure Data Factory
- Data flows allow data engineers to develop graphical data transformation logic without writing code
- The resulting data flows are executed as activities within Azure Data Factory pipelines that use scaled-out **Spark clusters**
- Data flow activities can be operationalized via existing Data Factory scheduling, control, flow, and monitoring capabilities



Azure Docs Link: [Concepts- Dataflow Overview]

ADFv2 : Key Concepts (6/6)

Integration Runtime

The Integration Runtime (IR) is the compute infrastructure used by Azure Data Factory to provide the following data integration capabilities across different network environments:

Data Flow: Execute a Data Flow in managed Azure compute environment.

Data movement: Copy data across data stores in public network and data stores in private network (on-premises or virtual private network). It provides support for built-in connectors, format conversion, column mapping, and performant and scalable data transfer.

Activity dispatch: Dispatch and monitor transformation activities running on a variety of compute services such as Azure Databricks, Azure HDInsight, Azure Machine Learning, Azure SQL Database, SQL Server, and more.

SSIS package execution: Natively execute SQL Server Integration Services (SSIS) packages in a managed Azure compute environment.

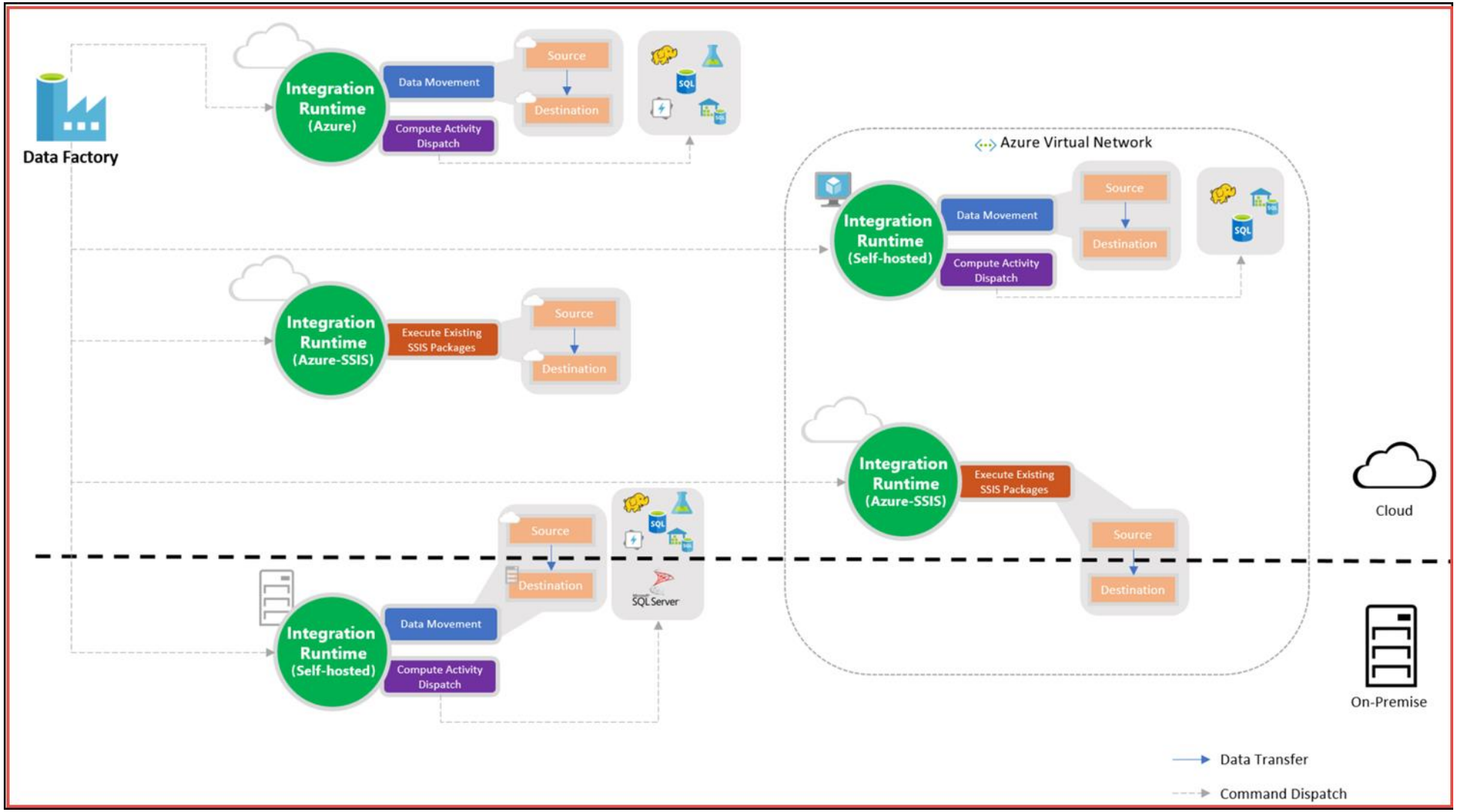
Integration Runtime Types

IR type	Public network	Private network
Azure	Data Flow	
	Data movement	
	Activity dispatch	
Self-hosted	Data movement	Data movement
	Activity dispatch	Activity dispatch
Azure-SSIS	SSIS package execution	SSIS package execution

Azure Docs Link: Concepts- Integration Runtime

ADFv2 : Key Concepts (6/6)

Integration Runtime



Azure Docs Link: Concepts- Integration Runtime

Create Azure Data Factory (Gen2)

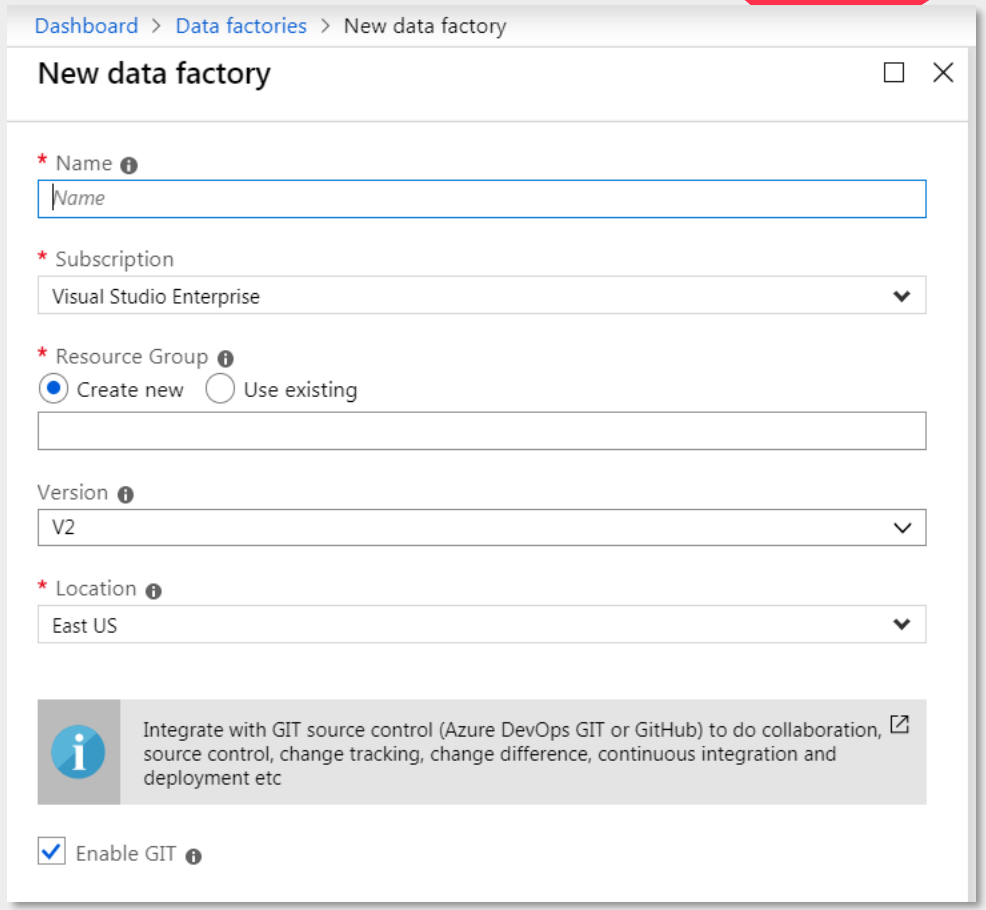
LAB

Select your existing resource group and select version v2.

The location throughout all services should be the same.

How are ADF Piplines created?

- Portal
- Visual Studio
- JSON (if you want to automate the generation)
- Powershell

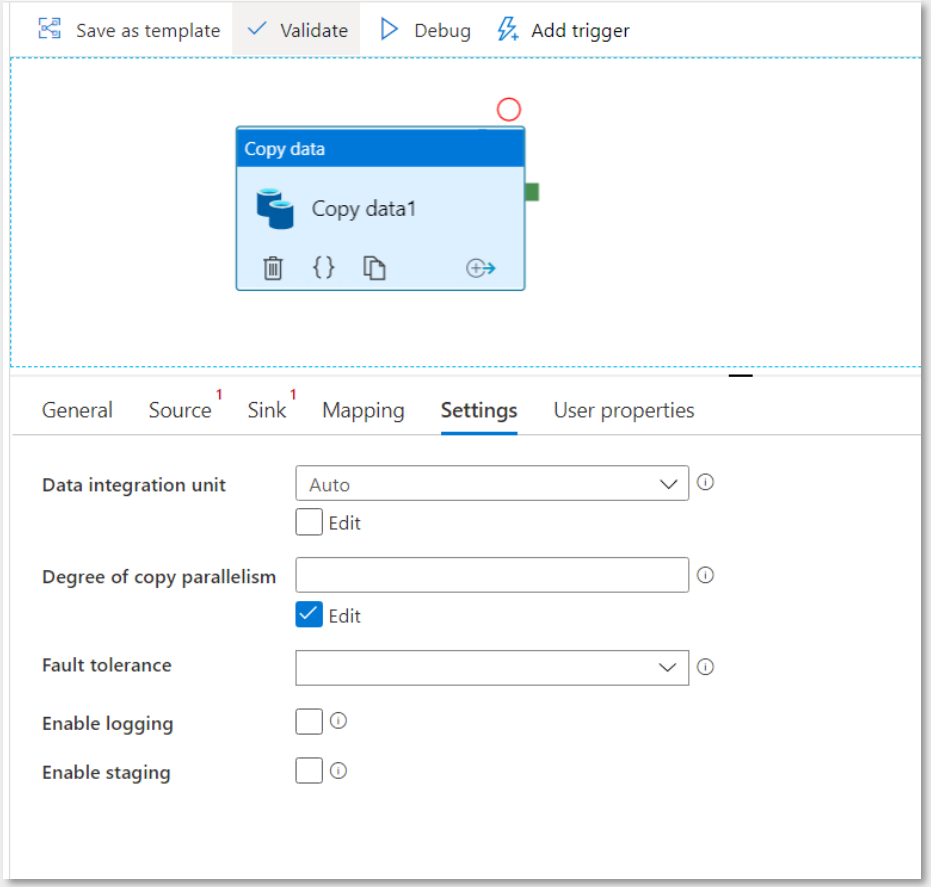


Copy Activity

LAB

Task to be completed:

- Using copy activity, copy the file from one folder to another folder.
- Copy data to Azure SQL
- Use different combination of delimitors like, esacpe sequence in the data, new line character in the data etc.
- Rejection Handling



PRO TIP:

- Use binary copy for just for moving file from one location to another
- When using Binary dataset, ADF does not parse file content but treat it as-is
- It makes the copy much faster

For-Each Activity

LAB

Data can be loaded in 20 parallel thread using For-Each activity

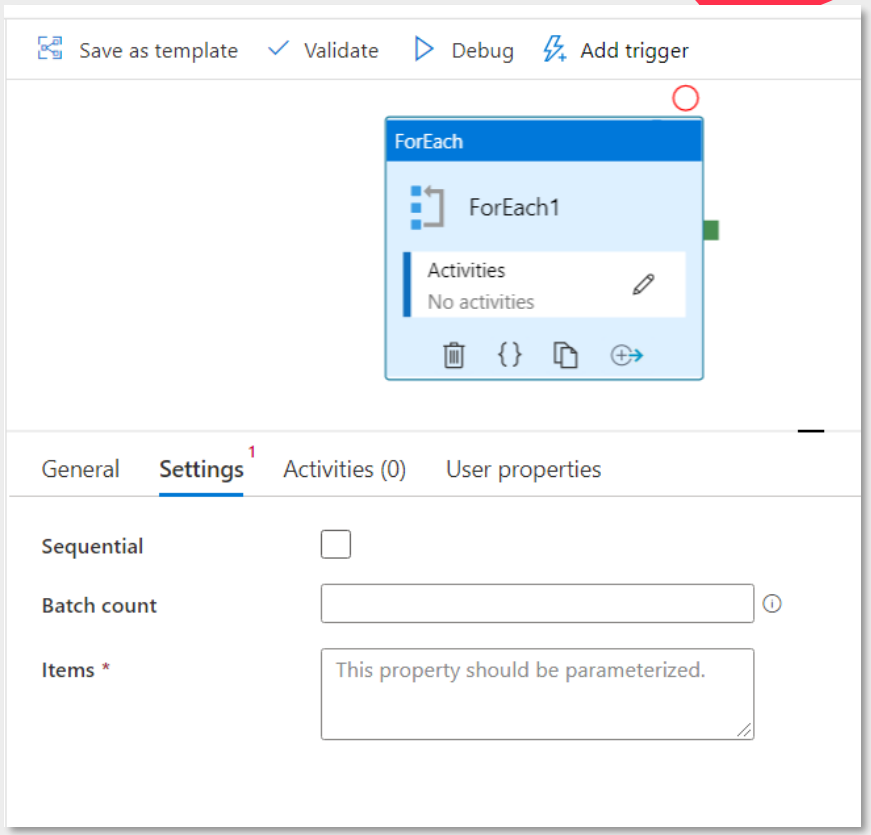
Task to be completed:

- Setup Lookup or Metadata Activity
- Load the five files in the table parallely

PRO TIP:

- There is auto retry for each task before it fails. It is very handy in scenarios where there is network lag or database is temporarily busy
- NOTE: Maximum parallel thread that can be executed by FOR-EACH is 20 (Link)

Ref: **Expressions and functions in Azure Data Factory**

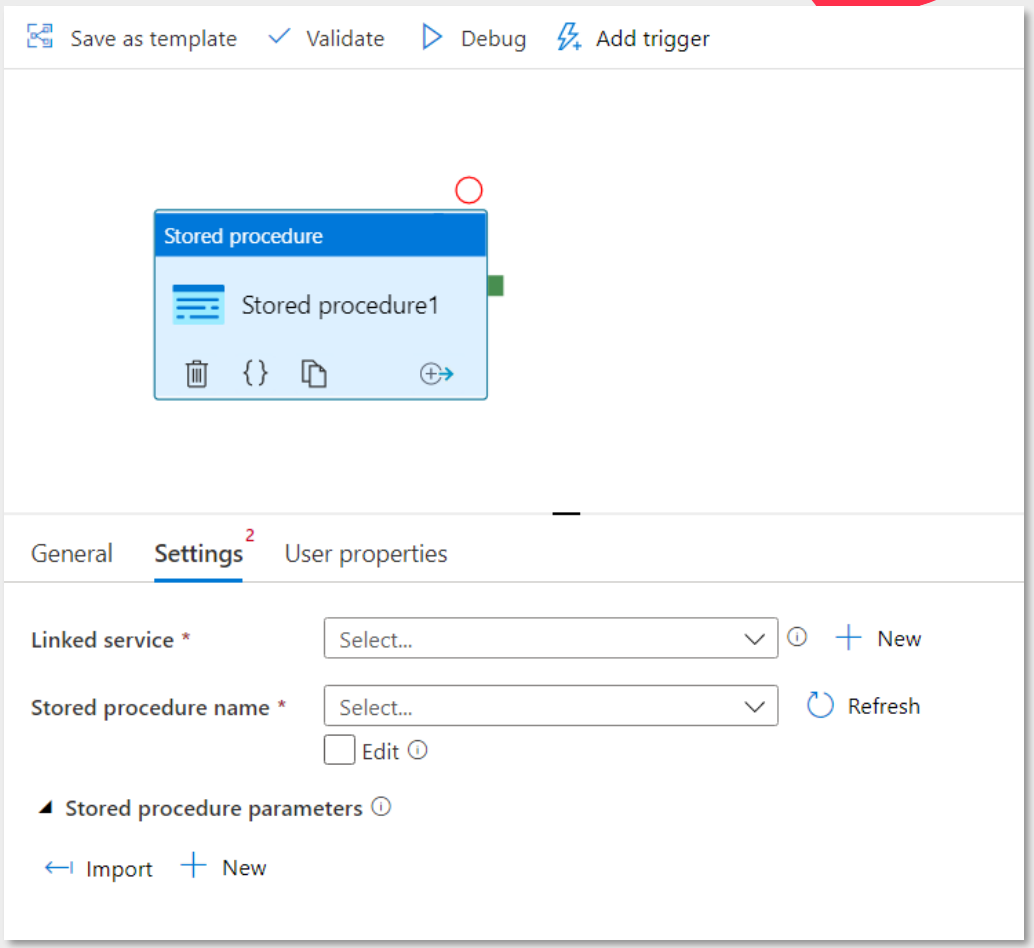


Stored Procedure Activity

LAB

Task to be completed:

- Call Stored procedure activity to update the loaded table.
- Add 0.1 to discount column

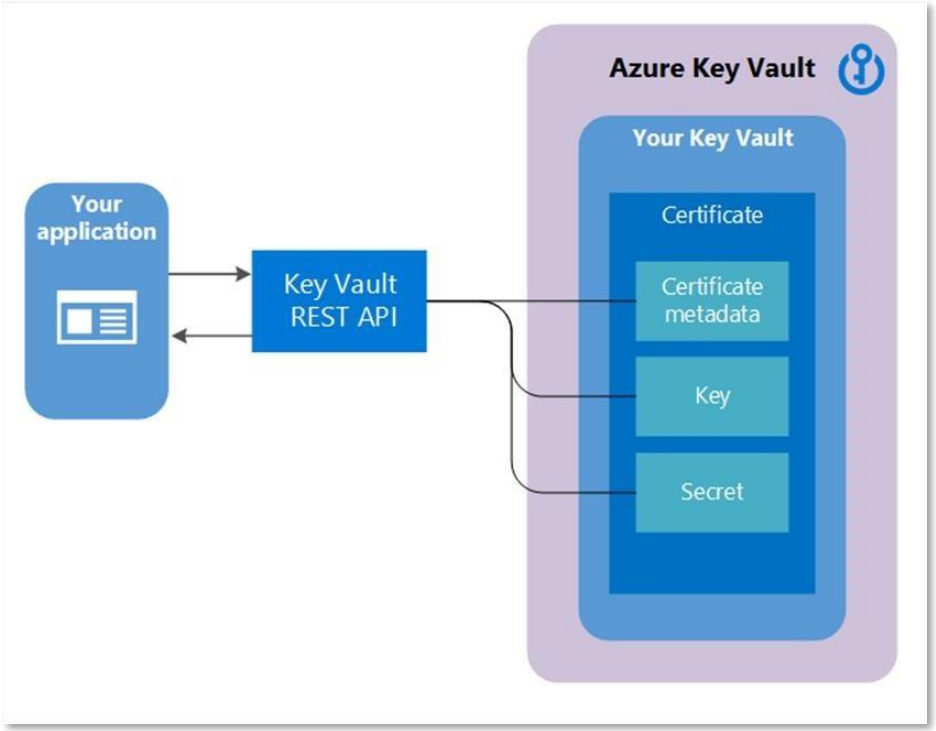


Azure Key Vault

Capgemini

Azure KeyVault

- Azure Key Vault is a cloud service for securely storing and accessing secrets.
- A secret is anything that you want to tightly control access to, such as API keys, passwords, certificates, or cryptographic keys.

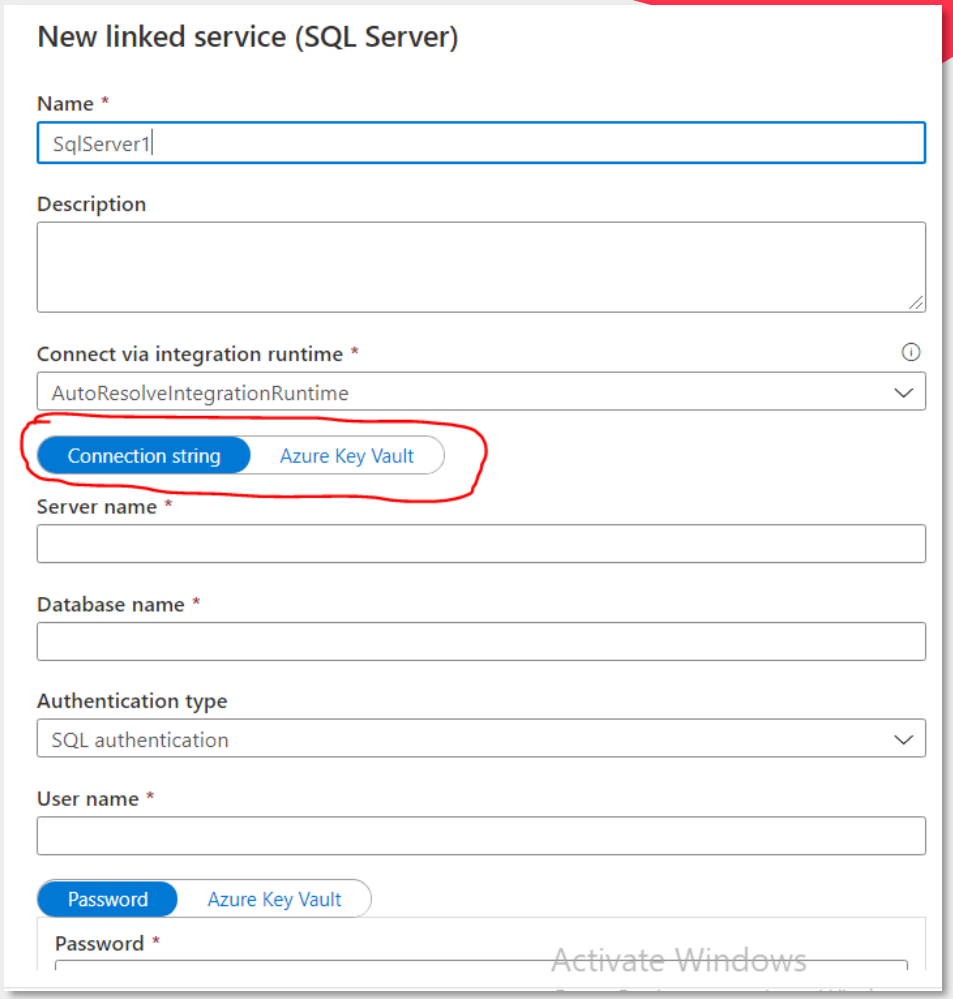


Using KeyVault in ADFv2

LAB

Task to be completed:

- Access Key Stored in KeyVault using ADF



Self Service LAB

- Difference Secanrio in delimited data, for eg delimiter in data, handling new line in data,
- Using different precedence constaints
- Binary Copy
- Stored Procedure Activity
- Using Lookup and Metadata activity
- Calling pipeline within pipelines
- For Loop – Parallelism
- Handling - Rejection of Records During Copy
- Fixed with copy with Databricks
- Expression and Variables and Parameters
- Naming Convension – Slide
- Batch Schedule
- Trigger – Lab
- DevOps with Azure
- Data Flow – Few Slides
- Self – Hosted Runtime : It is used on src and destination.

Reference

- Understanding block blobs, append blobs, and page blobs
- Expressions and functions in Azure Data Factory

About Capgemini

A global leader in consulting, technology services and digital transformation, Capgemini is at the forefront of innovation to address the entire breadth of clients' opportunities in the evolving world of cloud, digital and platforms. Building on its strong 50-year heritage and deep industry-specific expertise, Capgemini enables organizations to realize their business ambitions through an array of services from strategy to operations. Capgemini is driven by the conviction that the business value of technology comes from and through people. It is a multicultural company of over 200,000 team members in more than 40 countries. The Group reported 2018 global revenues of EUR 13.2 billion.

Learn more about us at

www.capgemini.com

People matter, results count.