Terms

Cloud computing

It's the delivery of computing services over the internet, which is otherwise known as the cloud. These services include servers, storage, databases, networking, software, analytics, and intelligence

Types of clouds

Public cloud

Services are offered over the public internet and available to anyone who wants to purchase them. Cloud resources, such as servers and storage are owned and operated by a third-party cloud service provider and delivered over the internet.

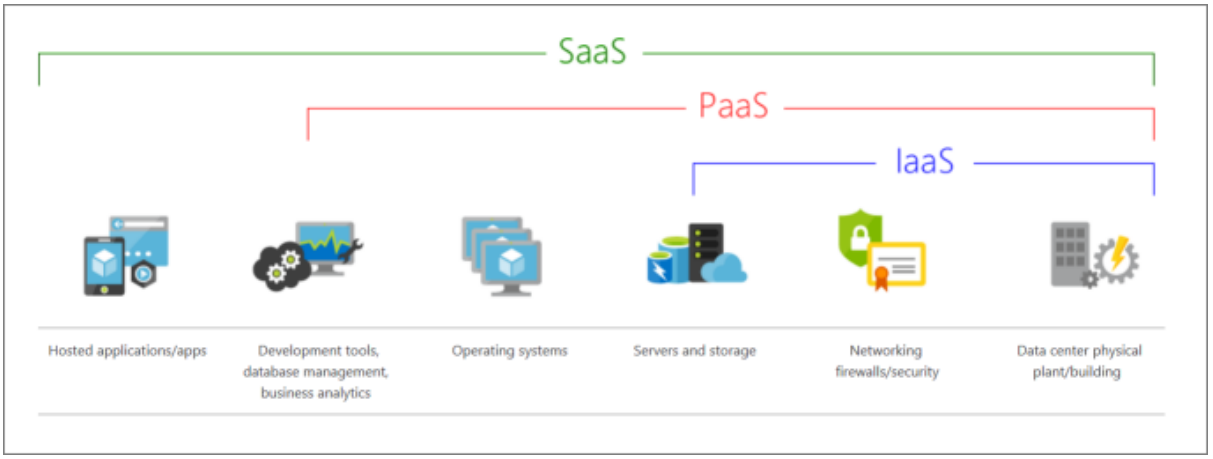
Private cloud

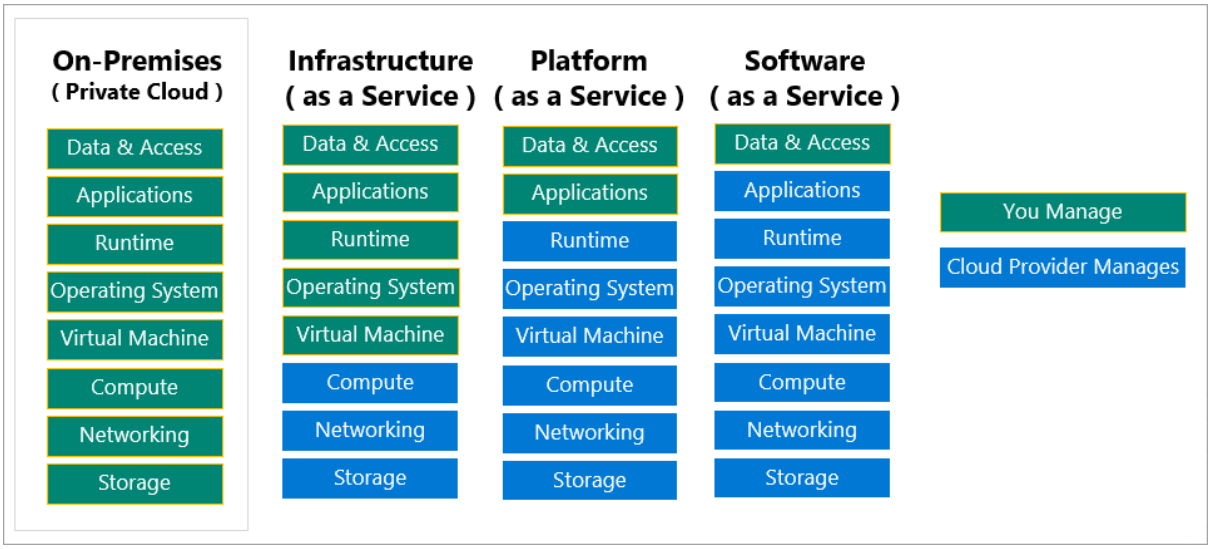
A private cloud consists of computing resources used exclusively by users from one business or organization. A private cloud can be physically located at your organization's on-site (on-premises) datacenter or it can be hosted by a third-party service provider.

Hybrid cloud

A hybrid cloud is a computing environment that combines a public cloud and a private cloud by allowing data and applications to be shared between them.

Cloud service models





IaaS (Infrastructure-as-a-Service)

This cloud service model is the closest to managing physical servers i.e. a cloud provider will keep the hardware up-to-date, but operating system maintenance and network configuration is up to you as the cloud tenant. For example, Azure virtual machines are fully operational virtual compute devices running in Microsoft datacenters. An advantage of this cloud service model is rapid deployment of new compute devices. Setting up a new virtual machine is considerably faster than procuring, installing, and configuring a physical server. IaaS is the most flexible category of cloud services. It aims to give you complete control over the hardware that runs your application. Instead of buying hardware, with IaaS you rent it.

PaaS (Platform-as-a-Service)

This cloud service model is a managed hosting environment. The cloud provider manages the virtual machines and networking resources and the cloud tenant deploys their applications into the managed hosting environment. For example, Azure App Services provides a managed hosting environment where developers can upload their web applications without having to worry about the physical hardware and software requirements.

SaaS (Software-as-a-Service)

In this cloud service model, the cloud provider manages all aspects of the application environment, such as virtual machines, networking resources, data storage and applications. The cloud tenant only needs to provide their data to the application managed by the cloud provider. For example, Microsoft Office 365 provides a fully working version of Microsoft Office that runs in the cloud. All you need to do is create your content and Office 365 takes care of everything else. SaaS is software that's centrally hosted and managed for you and

your users or customers. Usually one version of the application is used for all customers, and it's licensed through a monthly or annual subscription.

Azure Active Directory

Microsoft Windows Azure Active Directory (Windows Azure AD or Azure AD) is a cloud service that provides administrators with the ability to manage end-user identities and access privileges. Its services include core directory, access management and identity protection. The service gives administrators the freedom to choose which information will stay in the cloud, who can manage or use the information, which services or applications can access the information, and which end users can have access. Azure AD is the single and universal cloud-based identity and access management platform. Every organization will have an Azure AD or AD which helps employees to sign in and access various resources within the organization.

Subscriptions

A subscription groups together user accounts and the resources that have been created by those user accounts. For each subscription, there are limits or quotas on the amount of resources that you can create and use. Organizations can use subscriptions to manage costs and the resources that are created by users, teams, or projects.

Resources

Resources are instances of services that you create like virtual machines, storage, or SQL databases. A manageable item that's available through Azure. Virtual machines (VMs), storage accounts, web apps, databases, and virtual networks are examples of resources.

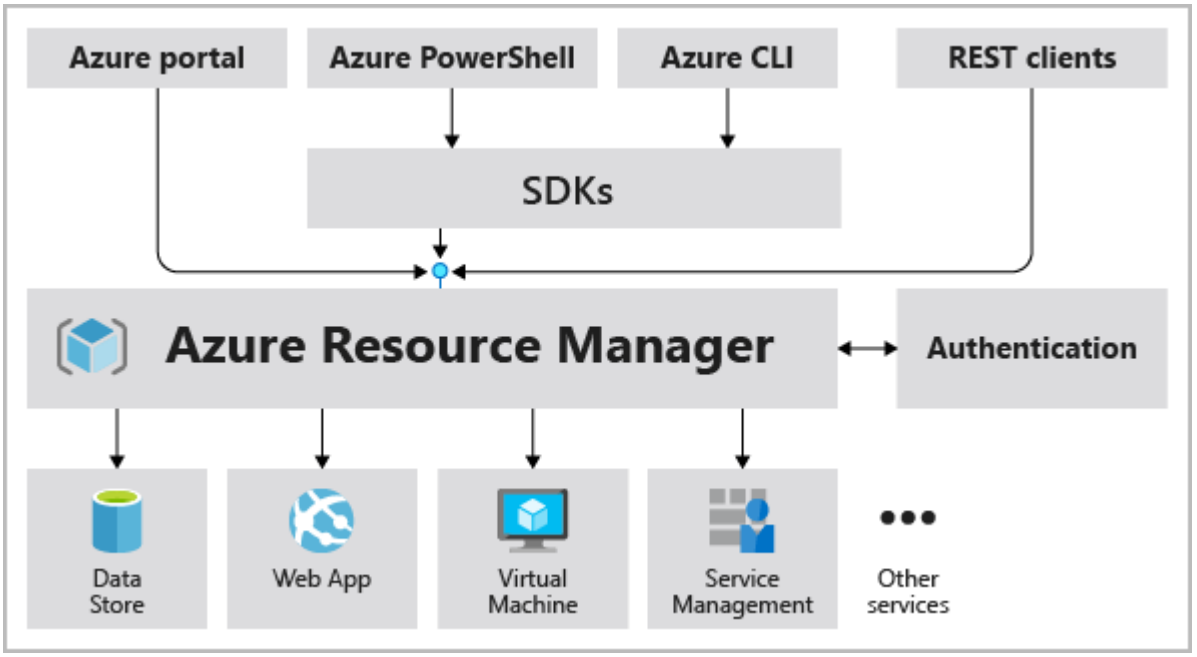
Resource groups

A container that holds related resources for an Azure solution. The resource group includes 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. A resource group is a logical container for resources deployed on Azure. These resources are anything you create in an Azure subscription like VMs, Azure Application Gateway instances, and Azure Cosmos DB instances. All resources must be in a resource group and a resource can only be a member of a single resource group. Many resources can be moved between resource groups with some services having specific limitations or requirements to move. Resource groups can't be nested. Before any resource can be provisioned, you need a resource group for it to be placed in.

Azure Resource Manager

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 account. You use management features like access control, locks, and tags to secure and organize your resources after deployment.

When a user sends a request from any of the Azure tools, APIs, or SDKs, Resource Manager receives the request. It authenticates and authorizes the request. Resource Manager sends the request to the Azure service, which takes the requested action. Because all requests are handled through the same API, you see consistent results and capabilities in all the different tools.



Authorisation

Resource groups are also a scope for applying role-based access control (RBAC) permissions. By applying RBAC permissions to a resource group, you can ease administration and limit access to allow only what's needed.

Workspace

A workspace defines the boundary for a set of related machine learning assets. You can use workspaces to group machine learning assets based on projects, deployment environments, teams, or some other organizing principle. The assets in a workspace include:

Compute targets for development, training, and deployment

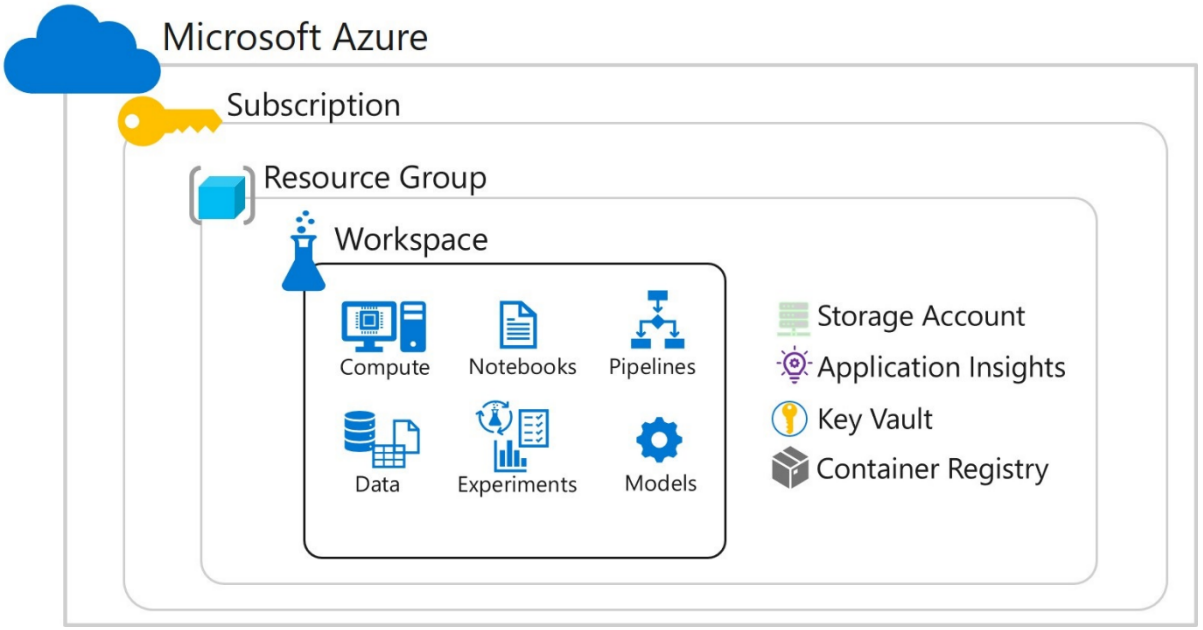
Data for experimentation and model training.

Notebooks containing shared code and documentation.

Experiments, including run history with logged metrics and outputs.

Pipelines that define orchestrated multi-step processes.

Models that you have trained



Management groups

These groups help you manage access, policy, and compliance for multiple subscriptions. All subscriptions in a management group automatically inherit the conditions applied to the management group.

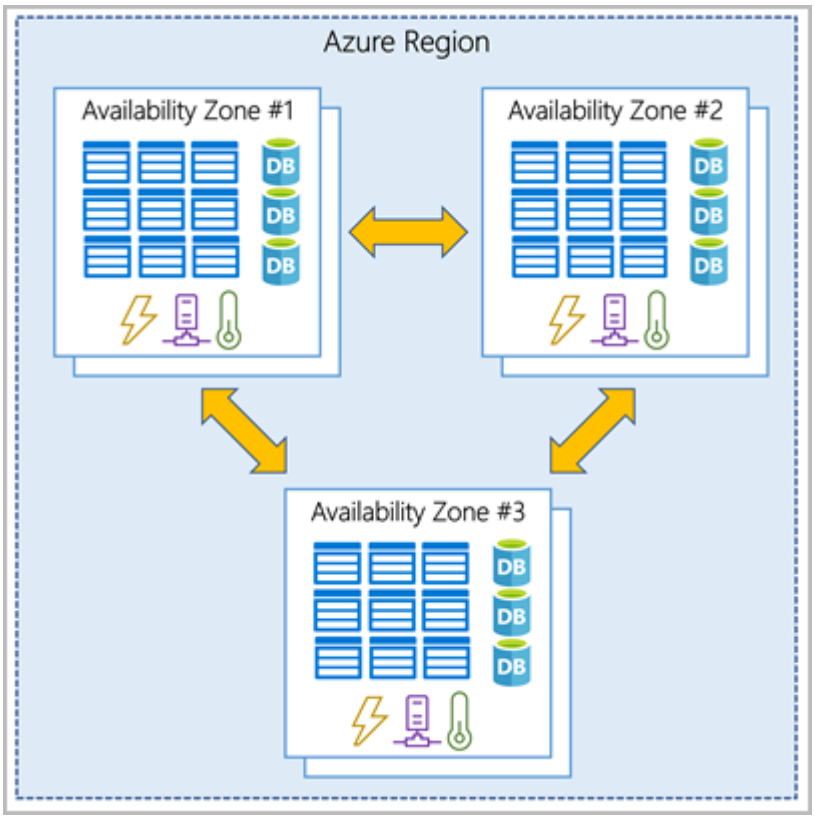
Azure regions

A region is a geographical area on the planet that contains at least one but potentially multiple datacenters that are nearby and networked together with a low-latency network Azure intelligently assigns and controls the resources within each region to ensure workloads are appropriately balanced. When you deploy a resource in Azure, you'll often need to choose the region where you want your resource deployed. Some services or VM features are only available in certain regions such as specific VM sizes or storage types.

There are also some global Azure services that don't require you to select a particular region such as Azure Active Directory, Azure Traffic Manager and Azure DNS.

Azure availability zones

Availability zones are physically separate datacenters within an Azure region. Each availability zone is made up of one or more datacenters equipped with independent power, cooling, and networking. An availability zone is set up to be an isolation boundary. If one zone goes down, the other continues working. Availability zones are connected through high-speed, private fiber-optic networks.



There's a minimum of three zones within a single region. It's possible that a large disaster could cause an outage big enough to affect even two datacenters.

Region pair

Each Azure region is always paired with another region within the same geography (such as US, Europe, or Asia) at least 300 miles away. This approach allows for the replication of resources such as VM storage across a geography that helps reduce the likelihood of interruptions because of events such as natural disasters, civil unrest, power outages, or physical network outages that affect both regions at once. If a region in a pair was affected by a natural disaster, for instance, services would automatically failover to the other region in its region pair.

Azure Machine Learning CLI Extension

The Azure command-line interface (CLI) is a cross-platform command-line tool for managing Azure resources. The Azure Machine Learning CLI extension is an additional package that provides commands for working with Azure Machine Learning.

Storage account

Storage account in Azure is a method of creating storage service for storing data in it. It contains all the all azure storage objects decided to single resource group. It contains Blob, queue, tables and files with disk images. It uniquely provides namespace and service access to functions of storage.

Compute instances

Development workstations that data scientists can use to work with data and models.

Compute Targets

In Azure Machine Learning, Compute Targets are physical or virtual computers on which experiments are run. A compute target is a designated compute resource or environment where you run your training script or host your service deployment

Types of compute

Azure Machine Learning supports multiple types of compute for experimentation and training. This enables you to select the most appropriate type of compute target for your particular needs.

Local compute - You can specify a local compute target for most processing tasks in Azure Machine Learning. This runs the experiment on the same compute target as the virtual machine such as an Azure Machine Learning compute instance on which you are running a notebook. Local compute is generally a great choice during development and testing with low to moderate volumes of data.

Compute clusters - For experiment workloads with high scalability requirements, you can use Azure Machine Learning compute clusters which are multi-node clusters of Virtual Machines that automatically scale up or down to meet demand. This is a cost-effective way to run experiments that need to handle large volumes of data or use parallel processing to distribute the workload and reduce the time it takes to run.

Attached compute - Links to other Azure compute resources such as Virtual Machines or Azure Databricks clusters. If you already use an Azure-based compute environment for data science such as a virtual machine or an Azure Databricks cluster, you can attach it to your

Azure Machine Learning workspace and use it as a compute target for certain types of workload.

Inference clusters - Deployment targets for predictive services that use your trained models. This kind of compute represents an Azure Kubernetes Service cluster and can only be used to deploy trained models as inferencing services.

Datastores

In Azure Machine Learning, datastores are abstractions for cloud data sources. They encapsulate the information required to connect to data sources. You can access datastores directly in code by using the Azure Machine Learning SDK and use it to upload or download data.

Azure Machine Learning supports the creation of datastores for multiple kinds of Azure data sources including Azure Storage (blob and file containers), Azure Data Lake store, Azure SQL Database, Azure Databricks file system (DBFS)

Every workspace has two built-in datastores (an Azure Storage blob container and an Azure Storage file container) that are used as system storage by Azure Machine Learning. There's also a third datastore that gets added to your workspace if you make use of the open datasets provided as samples

In most machine learning projects, you will likely need to work with data sources of your own, either because you need to store larger volumes of data than the built-in datastores support or because you need to integrate your machine learning solution with data from existing applications.

To add a datastore to your workspace, you can register it using the graphical interface in Azure Machine Learning studio or you can use the Azure Machine Learning SDK. You can view and manage datastores in Azure Machine Learning Studio or you can use the Azure Machine Learning SDK. The workspace always includes a default datastore and initially this is the built-in workspaceblobstore datastore

Dataset

Datasets are typically based on files in a datastore, though they can also be based on URLs and other sources. You can create the following types of dataset:

Tabular: The data is read from the dataset as a table. You should use this type of dataset when your data is consistently structured and you want to work with it in common tabular data structures such as Pandas dataframes.

File: The dataset presents a list of file paths that can be read as though from the file system. Use this type of dataset when your data is unstructured, or when you need to process the

data at the file level (for example - to train a convolutional neural network from a set of image files)

Environment

Python code runs in the context of a virtual environment that defines the version of the Python runtime to be used as well as the installed packages available to the code.

To improve portability, we usually create environments in docker containers that are in turn be hosted in compute targets, such as your development computer, virtual machines, or clusters in the cloud.

In general, Azure Machine Learning handles environment creation and package installation for you usually through the creation of Docker containers. You can specify the Conda or pip packages you need and have Azure Machine Learning create an environment for the experiment.

You can have Azure Machine Learning manage environment creation and package installation to define an environment and then register it for reuse. Alternatively, you can manage your own environments and register them. This makes it possible to define consistent and reusable runtime contexts for your experiments regardless of where the experiment script is run.

- Creating an environment from a specification file

You can use a Conda or pip specification file to define the packages required in a Python environment, and use it to create an Environment object.

- Creating an environment from an existing Conda environment

If you have an existing Conda environment defined on your workstation, you can use it to define an Azure Machine Learning environment:

- Creating an environment by specifying packages

You can define an environment by specifying the Conda and pip packages you need in a CondaDependencies object

Pipeline

In Azure Machine Learning, a pipeline is a workflow of machine learning tasks in which each task is implemented as a step.

Steps can be arranged sequentially or in parallel enabling you to build sophisticated flow logic to orchestrate machine learning operations. Each step can be run on a specific compute

target, making it possible to combine different types of processing as required to achieve an overall goal.

A pipeline can be executed as a process by running the pipeline as an experiment. Each step in the pipeline runs on its allocated compute target as part of the overall experiment run.

You can publish a pipeline as a REST endpoint, enabling client applications to initiate a pipeline run. You can also define a schedule for a pipeline, and have it run automatically at periodic intervals.

Azure Machine Learning SDK

While graphical interfaces like Azure Machine Learning studio make it easy to create and manage machine learning assets, it is often advantageous to use a code-based approach to managing resources.

You can use the Azure Machine Learning SDK to perform all of the tasks required to create and operate a machine learning solution in Azure. Rather than perform these tasks individually, you can use pipelines to orchestrate the steps required to prepare data, run training scripts, register models, and other tasks.

By writing scripts to create and manage resources, you can:

- Run machine learning operations from your preferred development environment.

- Automate asset creation and configuration to make it repeatable.

- Ensure consistency for resources that must be replicated in multiple environments (for example, development, test, and production)

- Incorporate machine learning asset configuration into developer operations (DevOps) workflows, such as continuous integration / continuous deployment (CI/CD) pipelines.

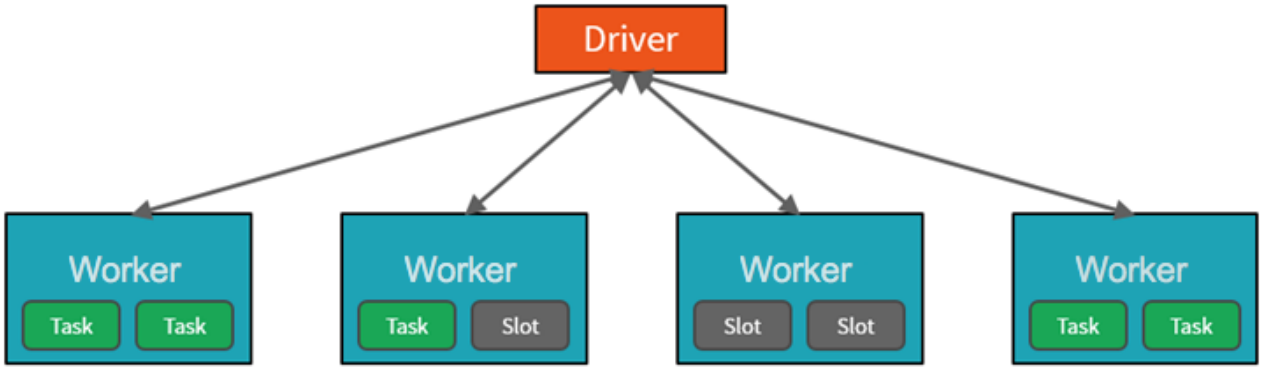
- Azure Machine Learning provides software development kits (SDKs) for Python and R, which you can use to create, manage, and use assets in an Azure Machine Learning workspace.

Azure Databricks

Azure Databricks provides a notebook-oriented Apache Spark as-a-service workspace environment. It is the most feature-rich hosted service available to run Spark workloads in Azure.

In Databricks, the notebook interface is the driver program. This driver program contains the main loop for the program and creates distributed datasets on the cluster, then applies

operations (transformations & actions) to those datasets. Driver programs access Apache Spark through a SparkSession object regardless of deployment location.



Some important points are

- The Driver is the JVM in which our application runs.

- Scaling vertically is limited to a finite amount of RAM, Threads and CPU speeds. Scaling horizontally means we can simply add new nodes to the cluster almost endlessly.

- We parallelize at two levels - the first level of parallelization is the Executor - a Java virtual machine running on a node, typically one instance per node. The second level of parallelization is the Slot - the number of which is determined by the number of cores and CPUs of each node. Each Executor has a number of Slots to which parallelized Tasks can be assigned to it by the Driver.

Driver create tasks for achieving parallelism and by creating tasks the driver can assign units of work to slots for parallel execution. Additionally, the driver must also decide how to partition the data so that it can be distributed for parallel processing. Once started, each task will fetch the partition of data assigned to it and work on it. Each parallelized action is referred to as a Job. The results of each Job is returned to the driver. Each Job is broken down into stages.