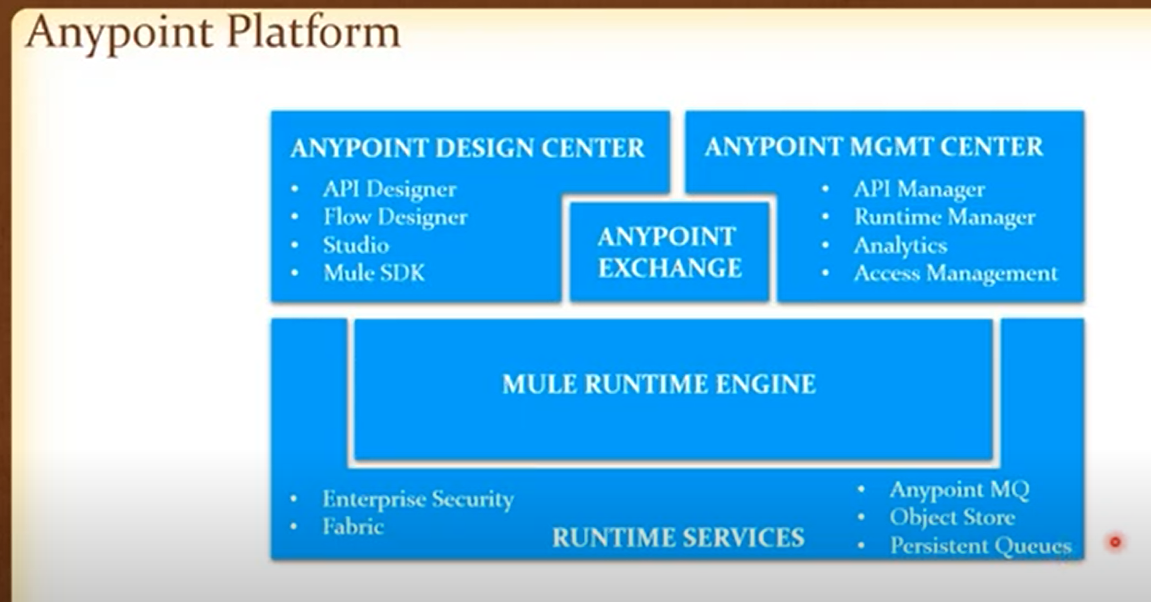
## Module 6: Runtime Manager



1. What is a Control plane?

* Control plane is the management centre.
* The components of the Anypoint Platform architecture that you use to design, deploy, and manage APIs and Mule applications are control plane.
* It contains product features and components that are part of the Anypoint Design Center, Anypoint Management Center, and Anypoint Exchange.
* Metadata- java utilization, jar libraries, Insights, analytic information related to API and runtime sit in the control plane region.
* CloudHub deployment log are stored in the control plane

1. What is a runtime plane?

* It is where mule runtime are located. It can be in laptop, bare metal or cloud like Azure
* Application are deployed and processed in the runtime plane
* The components of the Anypoint Platform architecture to which your APIs and Mule applications are deployed and made available to your users. The runtime plane includes the Mule runtime server and supporting services.
* All data payload are processed in runtime plane. Cloudhub is included in the runtime.

1. Features of the control plane?

The features are :

| **Component** |
| --- |
| API Designer |
| Anypoint Studio |
| Flow Designer |
| Anypoint Exchange |
| Anypoint API Community Manager |
| Anypoint Partner Manager |
| Access management |
| Analytics |
| Anypoint API Manager |
| Anypoint Runtime Manager |
| Anypoint Monitoring |
| Secrets manager |
| Anypoint Visualizer |
| Anypoint Security edge policies |
| Anypoint Security tokenization |
| CloudHub runtimes |
| Runtime Fabric |
| Customer-hosted runtimes |

1. Features of runtime plane?

//https://www.youtube.com/watch?v=sy5RS5wxkEA

1. What is network latency?

It is a delay in communication over a network.

It is the amount of time it takes for a packet of data to be captured, transmitted, processed through multiple devices, then received at its destination and decoded

1. What is a throughput?

**Throughput** is a measure of how many units of information a system can process in a given amount of time

### CloudHub Deployment

1. What is CloudHub?

CloudHub is an Integration Platform as a Service (iPaaS). It enables you to deploy and run the application in the cloud via Runtime Manager. CloudHub is a scalable, multi-tenant, elastic, secure, and highly available iPaas.

Features of a cloud hub:

* Zero down time
* Application isolation
* Default load balancer

1. How to deploy an application to CloudHub?

Applications can be deployed into cloud hub in following ways:

* Anypoint Runtime Manager
* Anypoint Studio
* CloudHub CLI
* Flow Designer in Anypoint Design Center
* CloudHub API
* Mule Maven plugin

1. What happens behind the scenes when you deploy an application to CloudHub?

When we deploy an application in a cloud hub, it creates an instances called elastic cloud instance (ec2) where mule jar is deployed. Each application deployed in the cloud hub are isolated from one another. It means each application have its own runtime. To managed the ec2 instances there is a shared load balancer

1. Advantages of deploying applications to CloudHub?

The advantage of using a Cloud hub are:

* + Get continuous software updates, no complex hardware maintenance required. Fully Managed
  + Improve efficiency with multi-tenancy for applications and workers
  + Enjoy a globally distributed architecture with 99.99% uptime. Globally available
  + Leverage out-of-the-box cloud security and compliance
  + Control access based on complex organizational requirements
  + Default load balancer
  + Zero down deployment
  + Application islation

1. Disadvantages of CloudHub deployment model?

Disadvantages are:

* + - Does o-not allow domain project deployment
    - Cannot customize the worker size, v core size and memory size
    - Cost - the minimal deployment unit is 0.1 vCore, meaning generally you can't run more than 10 Mule apps per single CloudHub license- resources are not used

1. What is a Shared Load Balancer (SLB)?

A shared load balancer in CloudHub resides outside the client's VPC. It's a shared resource, shared between customers and common for a specific CloudHub region. SLB is used for load balance external APIs.

1. What are the characteristics of a SLB?

The shared load balancer provides

* + - basic functionality, such as TCP load balancing.
    - have lower rate limits that help ensure platform stability
    - If an application exceeds the rate limit for a shared load balancer, the load balancer returns a 503 Service Unavailable response

1. What is an EC2 instance?

An **EC2 instance** is a virtual server in Amazon's Elastic Compute Cloud (**EC2**) for running applications on the Amazon Web Services (**AWS**) infrastructure

1. What is AWS and how CloudHub is related to AWS?

Amazon Web Services (**AWS**) is comprehensive and broadly adopted cloud platform

Cloud hub uses AWS as platform to run Mule applications

1. What is zero downtime deployment?

Zero down­time deploy­ment is a deploy­ment method where your web­site or appli­ca­tion is nev­er down or in an unsta­ble state dur­ing the deploy­ment process

Until up all request are served

1. What is a Worker?

**Worker** in cloudhub is a dedicated instance of Mule that runs your integration application.

Workers has following characteristics:

* + 1. Capacity: Each worker has a specific amount of capacity to process data. You can select the size of your workers when configuring an application.
    2. Isolation: Each worker runs in a separate container from every other application.
    3. Manageability: Each worker is deployed and monitored independently.
    4. Locality: Each worker runs in a specific worker cloud, the US, EU, Asia-Pacific, etc.

1. What is an availability zone?

Amazon EC2 is hosted in multiple locations world-wide. These locations are composed of Regions, Availability Zones, Local Zones, and Wavelength Zones. Each Region is a separate geographic area. Availability Zones are multiple, isolated locations within each Region

An Availability Zone (AZ) is one or more discrete data centers with redundant power, networking, and connectivity in an AWS Region. AZ’s give customers the ability to operate production applications and databases that are more highly available, fault tolerant, and scalable than would be possible from a single data center. All AZ’s in an AWS Region are interconnected with high-bandwidth, low-latency networking, over fully redundant, dedicated metro fiber providing high-throughput, low-latency networking between AZ’s. All traffic between AZ’s is encrypted.

1. What is a region w.r.t AWS data centers?

Each region has three data center

1. What is a \*\*vCore\*\* while deploying applications to CloudHub?

A unit of compute capacity for processing on CloudHub, which is equal to one virtual core. Up to ten **Mule** Applications can be deployed for every **VCore** purchased

1. What is application isolation?

Application isolation means each deployed application has its own resources and own mule runtime engine

1. What is vertical scalability?

Vertical scalability means increasing the size of the memory or CPU processing

1. How do you achieve vertical scalability?

It is achieved by increasing the number of vCores of each worker

// What is Scalability ? How can we achieve Scalability in MuleSoft Applications

<https://www.linkedin.com/pulse/what-scalability-how-can-we-achieve-mulesoft-sravan-lingam>

1. What is horizontal scalability?

Horizontal Scaling means adding a greater number of machines into our pool of resource

1. How do you achieve horizontal scalability?

In MuleSoft, Horizontal scaling can be achieved by increasing number of workers in CloudHub. Horizontal Scaling is divided into 2 types

Scale Up: Adding the more machines into your pool of resources.

Scale Down: Removing the machine from your pool of resources.

1. What is high availability?

The measure of a system’s ability to remain accessible in the event of a system component failure.

1. How can we achieve high availability in CloudHub deployment?

 HA is implemented by building in multiple levels of fault tolerance and/or load balancing capabilities into a system. To achieve HS application can be deployed in cluster or server groups so services is available even when one of the server is down

1. How can we achieve disaster recovery in CloudHub deployment?

Disaster recovery is a process by which a system is restored to a previous acceptable state, after a natural (flooding, tornadoes, earthquakes, fires, etc.) or man-made (power failures, server failures, misconfigurations, etc.) disaster

To achieve DR application can ne deployed in cluster or server groups in a distributed environment.

https://docs.mulesoft.com/runtime-manager/cloudhub-hadr

1. When should customers choose CloudHub deployment model?

CloudHub provides Fully-managed & highly available service . It is available globally in multiple regions. Also it provides secure and out of the box compliance environment in the cloud.

It is suitable for Experience APIs, Process APIs or Cloud to Cloud Integrations and Securely exposing APIs for Enterprise Systems hosted on premise.

<https://medium.com/@pkurimella/mulesoft-deployment-strategies-rtf-vs-hybrid-vs-cloudhub-e5aa241c365f>

Don't have a wider IT presence in my organization to manage the integration platform and need an agile solution to quickly build and deploy

### CloudHub VPC

1. What is a VPC?

A Virtual Private Cloud (VPC) allows you to virtually create a private and isolated network in the cloud. Just as a virtual private network (VPN) provides secure data transfer over the public Internet, a VPC provides secure data transfer between a private enterprise and a public cloud provider. This ensures that each customer's data remains isolated from other customer's data, both in transit and inside the cloud provider's network. This isolation can be accomplished using security policies that require some – or all – of the following elements: private IP addressing, tunneling, encryption, or allocating a unique VLAN to each customer.

<https://docs.mulesoft.com/runtime-manager/virtual-private-cloud>

1. What is CloudHub VPC?
2. Why do you need a VPC?

* Create a virtual private cloud
* It is used to create an isolated and dedicated network specific to a customer
* To have VPC user must create a separate account and different n/w for a cloud hub (IPSec)
* It creates a dedicated resources to the user
* Provides DLB
* VPN

1. Advantages of having a VPC?

* Security
* Vanity url
* Provides the ability to block certain (implementation) URL for security reasons like block customer to access the implementation url and redirect to proxy url
* Maintenance according to the customer choice
* Provide flexibility to the business
* DLB is available

1. Disadvantages of having a VPC?

Cost is high to operate

1. How do you configure a VPC?

<https://docs.mulesoft.com/runtime-manager/vpc-tutorial>

1. What is Dedication Load Balancer?
2. What are the features or characteristics of DLB?
3. Differences between SLB and DLB?
4. Advantages of having a DLB?
5. What are the ways that we can create a secure connection between CloudHub worker and customers' networks?
   * + - IPsec tunnel
6. What is an IPsec tunnel?

* IPsec is a group of protocols that are used together to set up [encrypted](https://www.cloudflare.com/learning/ssl/what-is-encryption/) connections between devices. It helps keep data sent over public networks secure. IPsec is often used to set up [VPNs](https://www.cloudflare.com/learning/access-management/what-is-a-vpn/), and it works by encrypting [IP](https://www.cloudflare.com/learning/network-layer/internet-protocol/) packets, along with authenticating the source where the packets come from.
* Within the term "IPsec," "IP" stands for "Internet Protocol" and "sec" for "secure."
* IPSec tunnel mode is the **default mode**. With tunnel mode, the entire original IP packet is protected by IPSec. This means IPSec wraps the original packet, encrypts it, adds a new IP header and sends it to the other side of the VPN tunnel (IPSec peer).

<https://www.youtube.com/watch?v=z-xGhij5MME>

1. What is direct connect?

Direct Connect is a cloud service from AWS that allows a dedicated network connection to be established between AWS and customers’ premises without using the Internet. Private connectivity can be established between AWS and an external location (e.g., data center, office, colocation environment).

1. What is VPC peering?

A VPC peering connection is a networking connection between two VPCs that enables you to route traffic between them using private IPv4 addresses or IPv6 addresses.

1. What is a vanity URL?

A **vanity URL** is a descriptive, memorable and pronounceable **URL** usually used to redirect **URLs** from one location to another. In the simplest terms, a **vanity URL** is a long **URL** that has been converted into a customized short link.

In the Domain Name System, a vanity domain is a domain name whose purpose is to express the individuality of the person on whose behalf it is registered. This contrasts with domain names which resolve to an organisation or a service that organisation offers.

1. How do you implement a two-way SSL using CloudHub VPC?

<https://help.mulesoft.com/s/article/One-way-and-2-way-SSL-Support-for-Cloudhub-Applications-Through-SLB-DLB>

1. When should a customer choose a CloudHub VPC deployment model?

* Private data centre. Need vanity url, secure data, Maintenance according to the customer choice
* Provide flexibility to the business

1. What are the firewall rules of VPC?

Four firewall rules are created by default—

* two rules that allow your private address space to connect through ports 8091 and 8092,
* two rules that whitelist external requests proxied by CloudHub’s shared load balancer through ports 8081 and 8082.
* All traffic that is not whitelisted in a firewall rule is blocked by default.

https://docs.mulesoft.com/runtime-manager/vpc-firewall-rules-concept

### Hybrid deployment

1. What is the hybrid type of deployment model?

* This model has:

On-premise or cloud runtime plane

Anypoint platform control plane

* Domain project deployment

With the hybrid deployment option, you deploy your apps from the Runtime Manager cloud console to your Mule servers and use Runtime Manager to manage them. This option provides you with flexibility and control over your on-premises security but requires you to provide the hosting infrastructure.  
To use the hybrid option, you first register your Mule servers with the Runtime Manager agent. Then, from Runtime Manager, you can optionally add those servers to server groups or clusters to provide high availability. Finally, you deploy your applications from Runtime Manager to either a server, server group, or cluster.  
Here the control plane is managed by mulesoft and runtime plane is managed by customer.

1. What are the types of hybrid deployment model?

The types of hybrid deployment model are

* Standalone server
* Cluster based: A cluster is a set of up to eight servers that act as a single deployment target and high-availability processing unit

application instances in a cluster are aware of each other, share common information, and synchronize statuses

If one server fails, another server takes over processing applications

A cluster can run multiple applications

* Server group

1. Advantages of hybrid deployment model?

* Key advantage of such a solution is that there are **no core limitations** per application; having just as many cores as our server has, we can deploy as many applications as Mule Server can handle, rapidly reducing licensing cost in comparison to Cloudhub
* This option provides flexibility and control over on-premises security
* provide high availability: Clustered Mule instances have distributed shared memory, which provides persistent VM queues, transactions, and cluster-wide data storage
* able to deploy domain project
* reduce licensing

1. Disadvantages of hybrid deployment model?

* requires the customer to provide the hosting infrastructure.
* it is necessary to see the release notes of the Mule (standalone) runtime engine to determine the minimum JDK supported version
* **Reduced out-of-the-box monitoring**
  + **Load balancing is not** provided for hybrid deployments.
  + The Runtime Manager **logging feature is not** available for hybrid deployments
  + Alerts feature not available
* Although Clustering Mule Runtime engines provide high availability it is still dependent on nodes availability and # of instances of mule apps running are dependent on # of clusters / servers.
* Auto-recovery of node is not available.
* Each runtime is limited to a specific version and any upgrades would require downtime.
* Application isolation is not available
* Need to manually update versions or patches
* Need an external load balancer
* Does not support mule 3 and mule 4 execution in same runtime/ resource

1. What is a standalone deployment?

It is a method to manually deploy a Mule application to on-premises Mule instance.

1. What is a cluster?

A cluster is a set of Mule runtime engines that acts as a unit. In other words, a cluster is a virtual server composed of multiple nodes (Mule runtime engines). The nodes in a cluster communicate and share information through a distributed shared memory grid. This means that the data is replicated across memory in different machines

A cluster is a set of up to eight servers that act as a single deployment target and high-availability processing unit. An application instances in a cluster are aware of each other, share common information, and synchronize statuses. If one server fails, another server takes over processing applications. A cluster can run multiple applications.

Deploying an application into a cluster is useful for achieving the following:

* High availability (HA): making your system continually available in the event that one or more servers in the cluster, or a data center, fails.
* Fault tolerance (FT): ensuring recovery from failure of an underlying component. Typically, the recovery is achieved through transaction rollback or compensating actions.
* Scaling: ensuring that your application can scale horizontally to meet increased demand.

1. What is a server-group?

A server group is a set of servers that act as a single deployment target for applications so that you don’t have to deploy applications to each server individually.

Deploying applications to servers in server groups provides redundancy so you can restore applications more seamlessly and quickly, with less downtime.

Unlike clusters, application instances in a server group run in isolation from the application instances running on the other servers in the group.

1. What is a hazelcast?

Hazelcast is a distributed In-Memory Data Grid platform for Java. The architecture supports high scalability and data distribution in a clustered environment. It supports the auto-discovery of nodes and intelligent synchronization

* It is a caching mechanism for mule runtime.
* It is a distribution memory grid that come along with standalone
* Hazelcast also offers an easy way to pass data or share state from one service to another
* provides the capability to monitor and alert on storage and use of data

1. How does node replication happen in a hybrid-based deployment model?

* When we go with cluster or server group deployment we can have node replication in hybrid

1. What is the use of node replication?

* High availability
* Reduces the workload on single node

1. What is distributed memory in a cluster-based deployment and how it is useful?
2. What is unicast in a cluster and when should customers choose unicast

A unicast cluster requires that we configure the IP addresses of the nodes in the cluster.

1. What is multicast in a cluster and when should customers choose multicast?

A multicast cluster comprises servers that automatically detect each other. Servers that are part of a multicast cluster must be on the same network segment.

One advantage of multicast clusters is that the server status doesn’t need to be Running to configure it as a node in a cluster. Another is that you can add nodes to the cluster dynamically without restarting the cluster.

1. How do you achieve High availability using a Cluster based deployment model?

https://docs.mulesoft.com/mule-runtime/4.3/mule-high-availability-ha-clusters

High Availability is a method of designing a computer system to prevent any downtime for the applications that run on it. Some systems use multiple servers so that if one server experiences downtime, the application can continue to run smoothly on the others, without interrupting service for the application’s end users

By default, clustering Mule runtime engines ensures high system availability. If a Mule runtime engine node becomes unavailable due to failure or planned downtime, another node in the cluster can assume the workload and continue to process existing events and messages. The following figure illustrates the processing of incoming messages by a cluster of two nodes. Notice that the processing load is balanced across nodes: Node 1 processes message 1 while Node 2 simultaneously processes message 2.

If one node fails, the other available nodes pick up the work of the failing node. As shown in the following figure, if Node 2 fails, Node 1 processes both message 1 and message 2.

Because all nodes in a cluster of Mule runtime engines process messages simultaneously, clusters can also improve performance and scalability. Compared to a single node instance, clusters can support more users or improve application performance by sharing the workload across multiple nodes or by adding nodes to the cluster.

1. When should a customer choose a standalone based hybrid deployment model?

* When customer want to deploy a domain project in one runtime

1. When should a customer choose a cluster-based hybrid deployment model?
2. When should a customer choose a server-group based hybrid deployment model?
3. How do you balance the load in hybrid?

* Have to configure external load balanger, Example nginx and apache

1. How can we achieve high availability in hybrid?
2. How can we achieve vertical and horizontal scalability in hybrid?

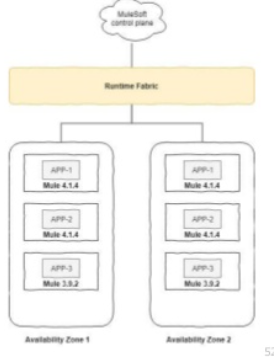
* Horizontal increasing the server node in cluster
* Vertical – increasing the memory size in .config file

### Runtime Fabric (RTF)

1. What is runtime fabric type of deployment?

Runtime Fabric (RTF) is a container based service that automates the deployment and orchestration of Mule applications and API gateways. RTF runs with in a cutomer managed infrastructure on AWS, Azure, VMs or bare metal servers

RTF is a deployment method:

* Mulesoft hosted control plane
* Customer hosted runtime plane
* 

1. When should a customer choose RTF deployment model?

It is used by the organization that have already setup an standalone infrastructure and need cloud hub features alongside then they for RTF

1. Advantages of using RTF?

* RTF offers abilities such as running **multiple versions** of Mule Runtimes on same servers,
* ability **to isolate applications** by running one Mule Runtime per application,
* Auto fail-over and CPU burst
* it **easier to switch between mule** versions thus giving an ability to do a phase based 3.x to 4.x migration on the same set of resources without changing the plumbing in place
* **allocate memory to low processing** application i.e flexibility to allocate memory
* **high availability** and reliability
* rolling upgrade, Zero downtime upgrades
* vertical and horizontal scaling
* load balancer included
* consistent deployment model for various cloud providers
* additional management center Ops- operation center: Allows to perform extensive monitoring
* additional security

1. Disadvantages of using RTF?

* No domain project store
* No object store
* Works on mule control plane

1. Differences between RTF and CloudHub deployment model?

* Cloud Hub
  + Mulesoft managed control plane and runtime plane
  + Vcore customization is not possible. Least is 0.1, 500 MB
* RTF
  + Mulesoft managed control plane and customer managed runtime plane
  + Vcore customization is possible. Least .02

1. Differences between RTF and Hybrid deployment model?

* Hybrid deployment
  1. Mule version update manually
  2. No default load balancer
  3. Do not support multiple version application
* RTF
  + - * 1. Mule version update manually
  1. has default load balancer
  2. support multiple version application

1. What is a Docker container?

A container is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another. A Docker container image is a lightweight, standalone, executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries and settings.

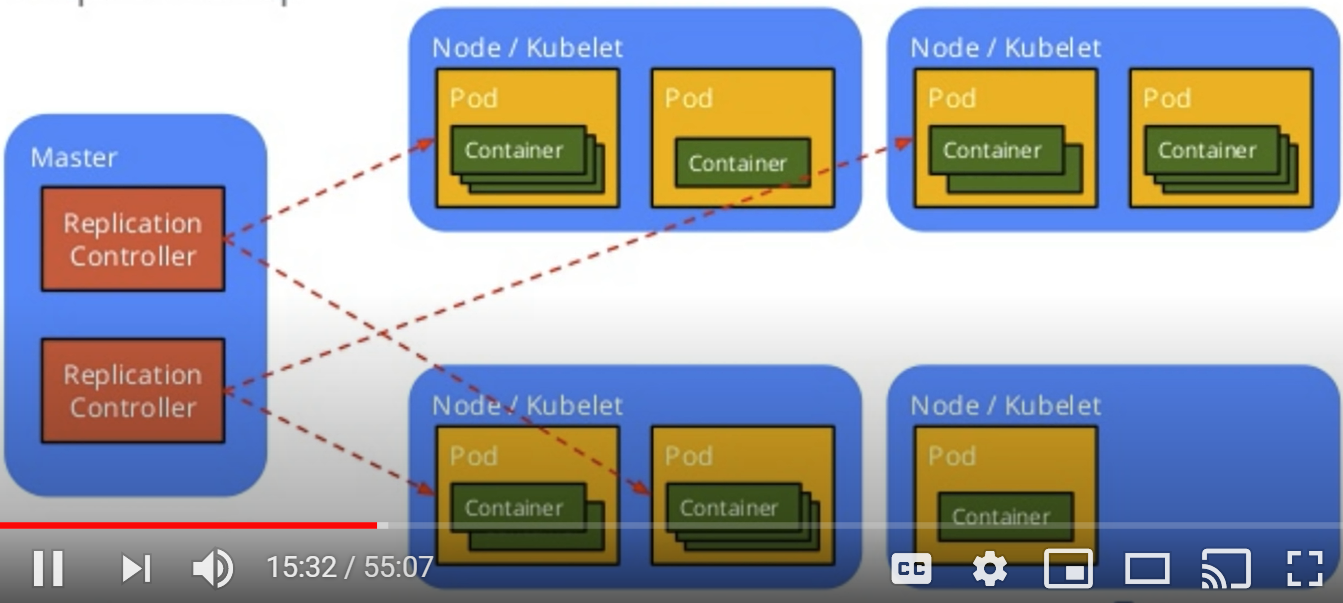
1. Differences between a Docker and a VM?

Docker is container based technology and containers are just user space of the operating system. At the low level, a container is just a set of processes that are isolated from the rest of the system, running from a distinct image that provides all files necessary to support the processes. It is built for running applications. In Docker, the containers running share the host OS kernel.

A Virtual Machine, on the other hand, is not based on container technology. They are made up of user space plus kernel space of an operating system. Under VMs, server hardware is virtualized. Each VM has Operating system (OS) & apps. It shares hardware resource from the host.

1. What is Kubernetes?

It is an open-source system for automating deployment, scaling, and management of containerized applications. It groups containers that make up an application into logical units for easy management and discovery.



1. Features of Kubernetes?
2. What is a controller?

In Kubernetes, controllers are control loops that watch the state of your [cluster](https://kubernetes.io/docs/reference/glossary/?all=true#term-cluster), then make or request changes where needed. Each controller tries to move the current cluster state closer to the desired state.

Controller: a VM dedicated for operating Runtime Fabric, including orchestration services, a distributed database, load balancing, and services that enable you to manage the cluster from Anypoint Platform.

The RTF controller is in fact a Kubernetes controller, so the orchestration, distributed database, and internal load-balancing capabilities are still being leveraged here. In addition, most control plane communication with AnyPoint Platform also happens here.  
The biggest consideration when deciding on the number and size of controllers is redundancy and load-balancing capacity. Requests to apps running on the worker nodes whether sourced from inside or outside the cluster will always go through the load-balancer on the controller nodes, ensuring the same clustering and zero-downtime upgrades, rollbacks, and scaling that CloudHub customers enjoy. However, keep in mind that if there are multiple controllers, an additional external load-balancer will be required for traffic from outside the cluster.  
The current maximum number of supported controller nodes is 5. Due to Kubernetes use of [raft](https://raft.github.io/) algorithm, it’s recommended to have an odd number of controllers.

The RTF worker node is where regular Mule applications will get deployed as pods. The more workers you have, the more highly-available and performant your cluster will be. However, the individual Mule application deployment configuration will decide on how much CPU, RAM, and number of replicas will be assigned to any one application.  
The current maximum number of supported worker nodes is 16.

1. What are the characteristics of a controller?

The characteristics of a controller are orchestration, distributed database, and internal load-balancing

1. What is a worker node?

It is the physical server or the VM that runs on the pod.

Worker: a VM dedicated for running Mule applications and API gateways. Mule applications and API proxies run on workers.

1. What is a POD?

It is a wrapper that contains one or more container. Example once container can have application and the other container can have a log

1. How do you achieve application isolation in RTF?

In RTF applications are deployed in the docker container which gives the application isolation. Runtin in a continaer

Every application that we deploy in RTF is deployed to a separate POD and container  
So that application isolation is achieved like cloudhub.

1. How do you achieve vertical and horizontal scalability in RTF?

Mule applications run on separate worker. This separation of responsibilities enables scaling of the worker nodes based on the number of Mule applications. It also enables scaling the controller nodes based on the frequency of deployments, changes in application state, and amount of inbound traffic.

1. How do you achieve high availability in RTF?

RTF uses cluster where applications are replicated. Thus it provides high availability

By deploying replicas of the mule apps, we can achieve high availability

Using docker and kubernetes.

1. How do you achieve zero-downtime in RTF?

* Whenever we deploy an updated app of an existing app onto RTF, a new POD created for this app by keeping the old one intact. This old one gets deprecated only when the new one is up and running. In this way 0 downtime is achieved by RTF.

1. What are the minimum requirements of controller and worker nodes?

Recommended number of controller nodes for development and production are as

* Production – 3
* Development – 1

Worker nodes

* Production – 3
* Development – 2

1. How do you balance the load in RTF?

RTF have internal load balancers within the controller nodes that directs the requests to the worker nodes. However with increase in number of controller nodes an external LB has to be configured for efficient performance

1. What control plane capabilities are available for RTF?

https://docs.mulesoft.com/runtime-fabric/1.7/architecture

### Private Cloud Edition (PCE)

1. What is PCE?

PCE stands for Private Cloud Edition, it is a deployment method in which both the control plane and runtime plane are hosted by the customers

In this the control plane components like design center and management are released into docker containers. Those docker containers are managed by the Kubernetes. The Runtime plane has the bare metal infrastructure. On top of it are the VMs and all the applications are deployed in those VM. The control plane uses the Docker and Kubernetes while the runtime plane is the collection of the clusters

1. When should a customer choose a PCE deployment model?

If an organization have a strict compliance laws on data processing within their own data centers than PCE is suitable. In such case the organization won’t even allow the low of meta-data outside the premises. Examples industries such as banks, insurance corps, healthcare and government organization have scenarios where they want to more tightly manage their data

1. Advantages of PCE?

The advantages are:

* Control Plane and Runtime plane are secured by the organization
* No flow of metadata to the Mulesoft hosted cloud
* All the exchange assets are strictly remain on premise
* The use of Kubernetes provides high availability and scalability

1. Disadvantages of PCE?

The disadvantages are:

* Requires expertise to manage the control plane
* Sometimes its difficult to resolve issues
* Since customers have to maintain the infrastructure it is costly

1. What are the differences between RTF and PCE?

Differences are:

* In RTF has customer hosted runtime plane and Mule soft hosted control while in PCE customer manages both the control plane and runtime plane
* In RTF there is a external security feature but this is not available in PCE. Customer has to maintain the security

1. What control plane capabilities are available for PCE?

The control plane capabilities available are:

* Anypoint Runtime Manager
* Anypoint API Manager
* Anypoint Exchange
* Access Management
* Anypoint Design Center, excluding Flow Designer

1. What are the basic configurations of a PCE?

Mulesoft support two types of configurations:

* + - * 3- Node configuration
      * 6- node configuration
      1. 3- Node configuration

Three is the minimum number of nodes that can enable high-availability and failover. In this configuration, each node hosts the platform applications and services. You must configure a load balancer to use round-robin distribution of traffic among each of the three nodes

Each node also hosts an instance of the database and object store. Hosting the database and object store enables persistence but requires larger minimum disk and memory requirements. Although each node contains a database, only one database is used as the master. Applications on each node write to this database only. The other two database instances are hot standby instances of the master database that take over as the master database in case of failure

* + - 1. 6- node configuration

In the 6-node configuration, three nodes are dedicated to hosting platform applications and services. The other three nodes host the database and object store instances. Although each node contains a database, only one of the database nodes is used as the master. Applications on each node write to this database only. The other two database instances are hot standby instances of the master database that take over as the master database in case of failure

1. How do you install the runtime plane in PCE?
2. How do you install the control plane in PCE?
3. Does PCE support application isolation and why?

PCE supports application isolation. All the applications are deployed in a docker which is managed by Kubernetes.

1. How do you achieve vertical and horizontal scalability in PCE?

As PCE uses Kubernetes. To achieve scalability we can change the size of node or pods.

**Horizontal scaling** means raising the amount of your instance. This can be done by adding new nodes to a cluster/pool. Or adding new pods by raising the replica count

**Vertical scaling** means raising the resources (like CPU or memory) of each node in the cluster (or in a pool). This is rarely possible without creating a completely new node pool. When it comes to pods though, vertical scaling would mean to dynamically adjust the resource requests and limits based on the current application needs

1. How do you achieve zero-downtime in PCE?

PCE uses server groups or clusters for deployment. They provide a single deployment target for applications so that user don’t have to deploy applications to each server individually. Thus it provides redundancy of applications. So in case of one server failure still the request from the user can be completed by another backup server. Thus it achieve zero down time

Redeployment – transistion time -

1. How do you balance the load in PCE?

In PCE user have to configure load balancer separately. The load balancer is places before the nodes so that all the incoming traffic are routed properly. Generally PCE uses round robin strategy for load balancing

### Pivotal Cloud Foundry (PCF)

1. What is PCF?

Pivotal cloud foundary hosting a control plane and Runtime plane

Pivotal cloud foundary can be hosted on customer Data center or AWS or GCP or on any cloud providers

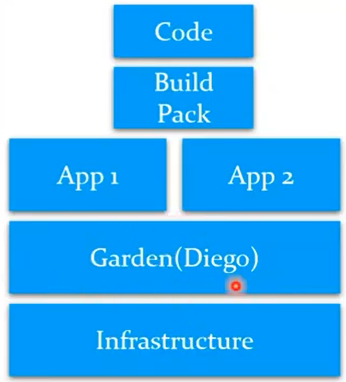
If we have an infrastuctutre On top of ‘Infrastructure’ ,PCF provides own  container engine which is  like Garden or Docker or Kubernetes

On top of container engine ,we  deploy our containerized applications.

PCF determines mule applications with the help of ‘Build Pack’

What usaually happens is when ever you create a code ,the code has some thing called as build pack

Job of ‘Build Pack’ is to determine what kind of application it is , mulesoft has created its PCF build packs , using those build packs PCF determines those are mule applications and they would require a mule runtime and accordingly it will help to deploy the application



1. When should a customer choose a PCF deployment model?
   * A typical **use** case for **PCF** is when companies want to deploy their applications on-premises for any reason (cost efficiency, flexibility, legal regulations, control over infrastructure, etc.). In this case they decide to **use PCF** as a leverage to build and operate their own (private) cloud offering.

* If customer organization has strict regulatory or compliance requirements
* If customer has PCF or cloud foundry in his landscape and want to have mule control plane/runtime hosted in the same cloud platform for better latency
* Compliance requirements: Government specific projects
* GDPR and Data residency requirements

1. Advantages of PCF?

* Latency: when your core systems are in one data center and you want your runtime to be in same data center
* Compliance requirements: Government specific projects GDPR and Data residency requirements

1. Disadvantages of PCF?

* High Availability: Customer need to configure
* Load Balancing: Customer need to configure
* Logging: No , need to leverage third party logging services like Splunk
* Monitoring: We can only trigger Alerts , can’t see utilization info like in Cloud Hub control plane
* Object Store: Customer need to configure
* Scheduling: No
* Security Updates: Manual

1. Explain the basic architecture of PCF.

* ·        Pivotal cloud foundary hosting a control plane and Runtime plane
* ·        Pivotal cloud foundary can be hosted on customer Data center or AWS or GCP or on any cloud providers
* ·        If we have an infrastuctutre On top of ‘Infrastructure’ ,PCF provides own  container engine which is  like Garden or Docker or Kubernetes
* ·        On top of container engine ,we  deploy our containerized applications.
* ·        PCF determines mule applications with the help of ‘Build Pack’
* ·        What usaually happens is when ever you create a code ,the code has some thing called as build pack
* ·        Job of ‘Build Pack’ is to determine what kind of application it is , mulesoft has created its PCF build packs , using those build packs PCF determines those are mule applications and they would require a mule runtime and accordingly it will help to deploy the application.
* ·        Like PCE ,Control Plane and Runtime Plane are hosted in customer’s data center
* ·        Unlike manual runtime installation in PCE, runtimes are provisioned automatically using PCF technology .
* ·        Unlike PCE, every application will get separate runtime in PCF
* Buildpacks provide framework and runtime support for apps.
* ·        Buildpacks typically examine your apps to determine what dependencies to download and how to configure the apps to communicate with bound services.
* When you push an app, Cloud Foundry automatically detects an appropriate buildpack for it. This buildpack is used to compile or prepare your app for launch.

1. What are the differences between RTF and PCF?

In RTF control plane is hosted on MuleSoft while in PCF control plane is hosted on customer premise

In RTF application are deployed on AWS or standalone while in PCF in addition to this it has a cloud foundry infrastructure in which applications are deployed

In RTF it has anypoint security component in control plane but lack in PCF

1. What are the differences between PCE and PCF?
   * PCE and PCF have similar functionality but in case of PCF it has a cloud foundry on which control plane and runtime lane are placed

* Instead of customer manually installing mule runtime plane ,the way runtimes are provisioned is automatic using PCF build packs technology .It uses PCF droplets , so for each application PCF droplet gets created which will have  a mule runtime .Applications will be deployed on to PCF droplets.
* Will get application isolation here as well

1. What control plane capabilities are available for PCF?
   * + - **Runtime Manager Service**:
2. What are droplets in PCF and how is it used?

Job of ‘Build Pack’ is to determine what kind of application it is , mulesoft has created its PCF build packs , using those build packs PCF determines those are mule applications and they would require a mule runtime and accordingly it will help to deploy the application.

1. How do you install the runtime plane in PCF?

Runtime plane gets generated automatically using PCF build packs technology

1. How do you install the control plane in PCF?

Install the Any point platform Service Brokers for PCF tile.

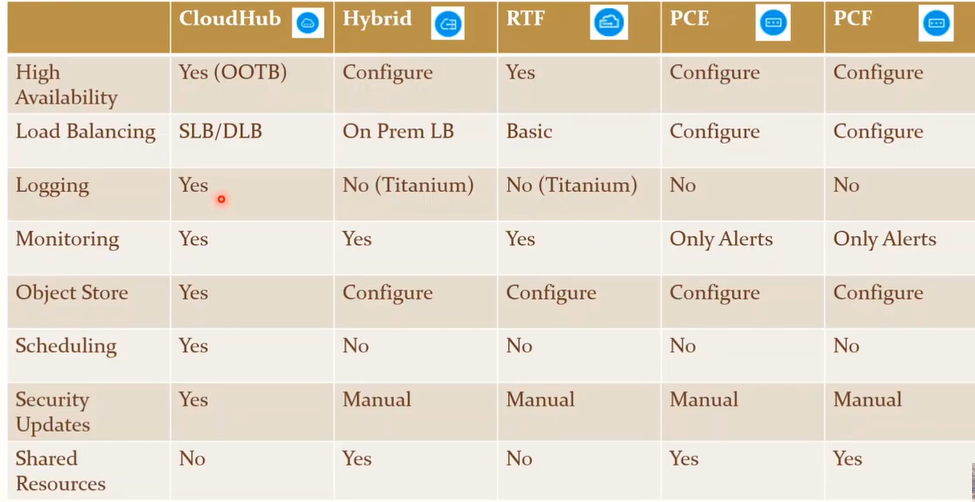
1. Does PCF support application isolation and why?

Yes , It supports application isolation , because for each application PCF droplet gets created which will have  a mule runtime .Applications will be deployed on to PCF droplets.

1. How do you achieve vertical and horizontal scalability in PCF?

Horizontally scale their Mule applications by increasing the number of replicas.

1. How do you achieve zero-downtime in PCF?
2. How do you balance the load in PCF?



<https://www.youtube.com/watch?v=NChZm78YVis&amp;t=507s>

### Runtime Behaviour / Thread pool

1. What is a thread?

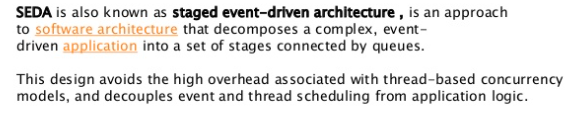
A **thread** of execution is the smallest sequence of programmed instructions that can be managed independently by a scheduler,

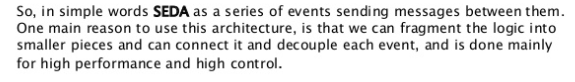
1. What is a thread pool?

A **thread pool** is a collection of worker **threads** that efficiently execute asynchronous callbacks on behalf of the application. The **thread pool** is primarily used to reduce the number of application **threads** and provide management of the worker **threads**.

1. What is staged event-driven architecture (SEDA)?

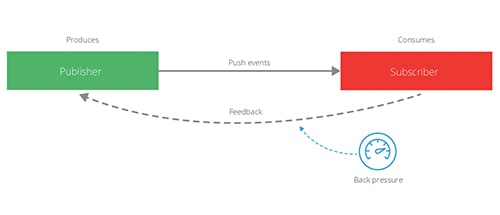
is designed to enable high concurrency of services. SEDA decomposes an application into network of stages separated by dynamic resource controllers to allow applications to adjust dynamically to changing load.





https://dzone.com/articles/mule-sub-flows-processing-strategy-and-one-way-end

1. What is reactive programming?



Mule 4 uses reactive programming to enhance scalability so our developers can deliver smooth experiences for their users.

 It is a programming paradigm that is asynchronous, non-blocking and declarative that enables working with stream of data rather than one single all in-memory collections

1. How reactive programming is used in mule?
2. What are the features of reactive programming?

* Responsive:  The system responds in a timely manner if at all possible. Responsive systems focus on providing rapid and consistent response times.
* Resilient: The system stays responsive in the face of failure. This applies not only to highly-available, mission critical systems. Any system that is not resilient will be unresponsive after a failure. Resilient is achieved by containment, isolation and replication.

* Elastic: The system stays responsive under varying workload and bottleneck. Reactive Systems can react to changes in the input rate by increasing or decreasing the resources allocated to service these inputs

* Message Driven: Reactive Systems rely on [asynchronous](https://www.reactivemanifesto.org/glossary#Asynchronous) [message-passing](https://www.reactivemanifesto.org/glossary#Message-Driven) to establish a boundary between components that ensures loose coupling, isolation and location transparency.

1. What is backpressure?

In order to allow the subscriber to consume the event stream without being overwhelmed by too many events, the subscriber must be able to apply back pressure, in other words signal to the producer to “please slow down.”

1. How backpressure is implemented in mule?

**Automatic back pressure in Mule 4**

**Manual back pressure in Mule 4**

1. What is automatic back-pressure?

Mule 4 applications are automatically configured so that the event source receives a back pressure signal when all threads are currently executing and no free threads remain in a required thread pool. In practical terms this will trigger the HTTP Listener, for example, to respond with a 503–“Server busy”, and the JMS Listener will not acknowledge receipt of a message. OutOfMemory errors are avoided as a result of this configuration.

1. What is manual backpressure?

Mule developers can also configure each event processor to signal back pressure to the event source through the “**maxConcurrency**” attribute. This configuration affects the number of events that can pas through the event processor per second.

1. How to configure a manual backpressure?

By setting the “**maxConcurrency**” attribute. Example batch and parallel for each

1. What is auto tuning in mule and how does it work?

Thread pools are not configurable at the level of Mule application. It will be decided by  
the runtime itself. Mule 4 calculates the sizing of thread pools dynamically and  
automatically, and in most scenarios the defaults are optimal. Under most  
circumstances, it not recommended from MuleSoft to change the default values.

<https://blogs.mulesoft.com/biz/mule/thread-management-auto-tuning-mule-4/>

1. What are the different processing types in Mule?

CPU lite, CPU intensive and blocking IO

1. What is CPU\_LITE processing type?
   * Used for the tasks that takes at most 10ms to execute.
   * No blocking IO operations should be executed here
   * Number of threads = (Thread Default size is ) 2 \* cores
   * Logger, set payload
2. Examples of connectors/processors which use CPU\_LITE processing type?

Logger, set payload, set variable

1. What is CPU\_INTENSIVE processing type?

For tasks that take more than 10ms to execute (duration is not enforced, but misclassifying tasks has bad consequences).

Typically for transformations, encrypt/decrypt, heavy computation, etc. like data weave and scripting

Non-blocking IO operations should be executed here.

Number of threads = (Thread Default size is ) 2 \* cores

1. Examples of connectors/processors which use CPU\_INTENSIVE processing type?

Transform component, Encrypt

1. What is the BLOCKING\_IO processing type?
   * All blocking IO operations should happen here.
   * Significantly larger than the other pools, as most threads here are expected to be in a blocked state.
   * Default size comes from a formula that considers the available memory, the default size of the streaming buffers and other concepts.
   * Database, I/O opertions components like FTP
2. Examples of connectors/processors which use BLOCKING\_IO processing type?

Database, I/O opertions components like FTP

1. What is the Proactor pattern?

Proactor is a design pattern for asynchronous execution. It segregates all the task that  
will be executed in the flow into respective categories and then assigns required thread  
pools to each of those categories

1. What is Grizzly thread pool?

**Http connector** uses separate thread pool called selector pool for its processing. It uses  
**Java Grizzly** libraries internally. These thread pools are called Grizzly thread pool

The minimum size of the shared Grizzly pool for the HTTP Listener is determined upon the deployment of the first app to the Mule runtime that uses an HTTP Listener. The size of the dedicated Grizzly for the HTTP Requester pool is determined upon deployment of each app that uses an HTTP Requester.

1. What is Grizzly Shared thread pool

Thread pool used by **Http Listener** component is Grizzly shared thread pool, this has to  
be **shared among** all the applications running on the same mule runtime

1. What is Grizzly Dedicated thread pool?

Thread pool used by **Http Request** component is Grizzly dedicated thread pool, there  
will be **dedicated thread** pool for each application running in the same mule runtime

1. What is Java NIO and how is it used in mule?

java NIO (New IO) is an alternative IO API for Java (from Java 1.4), meaning alternative to the standard [Java IO](http://tutorials.jenkov.com/java-io/index.html) API’s.

Java NIO and NIO2 libraries provides helps so that threads do not block waiting for IO  
intensive operations.  
NIO Selector thread pool is a separate reserved custom thread pool which deals with  
blocking IO operations. In case there is a blocking IO call, the thread which was executing  
the process is released immediately back to its own pool so that it can perform other  
operations, and the blocking operation is delegated to these selector pools (These  
threads belong to the OS kernel and schedule management is taken care of by the  
underlying OS, more the number of base cores better will be the processing).

1. What is the selector thread and how is it used in mule?

Selector thread pool is a separate reserved custom thread pool which deals with  
blocking IO operations. In case there is a blocking IO call, the thread which was executing  
the process is released immediately back to its own pool so that it can perform otheroperations, and the blocking operation is delegated to these selector pools (These  
threads belong to the OS kernel and schedule management is taken care of by the  
underlying OS, more the number of base cores better will be the processing

There is a special thread pool for the HTTP Listener. This is configured at the Mule runtime level and shared by all applications deployed to that runtime. There is also a special thread pool for the HTTP Requester. This is dedicated to the application that uses an HTTP Requester.

1. What is thread switching?

Thread switching is a type of context switching from one thread to another thread in the  
same process

1. Why is thread switching expensive?

Switching the CPU from one thread to another involves suspending the current thread,  
saving its state (e.g., registers), and then restoring the state of the thread being  
switched to. The thread switch actually completes at the moment a new program  
counter is loaded into PC; at that point, the CPU is no longer executing the thread  
switching code, it is executing code associated with the new thread.

1. How does mule try to avoid thread switching?

Due to optimizations regarding latency, thread switches are omitted when an IO or  
CPU\_INTENSIVE task is followed by a CPU\_LIGHT one. Reasoning behind this  
optimization is that executing said CPU\_LIGHT task is most likely cheaper than the  
thread switch.

1. What is the UBER thread pool?

In 4.3 unifed three thread pools from 4.1 and 4.2 (cpu\_lite, cpu\_intensive, blocking\_io)  
into a single thread pool. This helps us improve the Mule runtime’s auto-tuning feature  
and make better use of available resources. It is called UBER POOL

1. What is a pooling strategy?

Pooling strategy indicate the type of thread pool used in Mule runtime. In mule 4.3  
default pooling strategy is UBER in which there is a only one thread pool  
Another option for pooling strategy is DEDICATED thread pool in which there will be  
separate thread pool for each type of processing types  
We can configure pooling strategy in scheduler-conf.properties file

1. How does mule increase t
2. What is JVM tuning?

JVM tuning mainly involves optimizing the garbage collector for better collection  
performance so that applications running on VMs can have a larger throughput while  
using less memory and experiencing lower latency

1. What is Heap memory?

Heap memory is allocated to store objects and JRE classes. It can be modified by  
changing wrapper.java.maxmemory in wrapper.config file

1. What is metaspace memory?

Metaspace memory is the memory allocated to store metadata about the application  
the JVM is running. It contains class definitions, method definitions, and other  
information about the program. Metaspace size can be limited by using the  
MaxMetaspaceSize option inside wrapper.config file