**Module 6: Runtime Manager**

1. What is a Control plane?

Control Plane includes the components of Anypoint Platform which helps in creating, managing and deploying APIs and Applications. Control Plane components include Design Centre, Management Centre, Runtime Manager

1. What is a runtime plane?

Runtime plane is where the application is deployed. It is where Mulesoft Runtime is available.

1. Features of the control plane?

The control plane contains product features and components that are part of the Anypoint Design Centre, Anypoint Management Centre, and Anypoint Exchange.

Control Plane have data such as the API analytics information like how many APIs are called, how much time is taken to complete the transaction, JVM utilization, CPU utilization, Jar file, Logs etc.

Different control plane options available are Mulesoft hosted (US control plane, EU control Plane), Mulesoft Government Cloud and Customer Hosted Control Plane

1. Features of runtime plane?

The runtime engine includes Anypoint Security edge policies and tokenization, MQ, Object Store, and Connectors. To deploy your applications and APIs via runtime plane, there are three options: CloudHub, Anypoint Runtime Fabric, or on–premises Mule runtime engine instances.

1. What is network latency?

It is the time taken for data or request to go from source to destination. Latency closer to zero is better

1. What is a throughput?

Throughput refers how much data can be transferred from one location to another in a given amount of time

Throughput Formula (TH) = I(Inventory)/ T(Time)

**CloudHub Deployment**

1. What is CloudHub?

CloudHub is an Integration Platform as a Service (iPaaS) which enables to deploy and run the application in the cloud via Runtime Manager

1. How to deploy an application to CloudHub?

We can deploy application either from Anypoint Studio or from Runtime Manager.

Runtime Manager

* Sign in to Anypoint Platform.
* Select Runtime Manager.
* In the Applications page, upload jar file of the application
* Click Deploy application

Anypoint Studio

* Open application in Studio.
* In Package Explorer, right-click the project folder and select Anypoint Platform > Deploy to CloudHub.
* If this is our first time deploying from Studio, provide your Anypoint Platform credentials at the prompt.
* Studio stores our credentials for future use.
* We can manage these credentials by selecting Anypoint Studio > Preferences > Anypoint Studio > Authentication.

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

An application running in Cloudhub means it is running in AWS. So when a new application is deployed to cloudhub means, a new ec2 instance is spinned up which will have the cloudhub worker and mule runtime.

1. Advantages of deploying applications to CloudHub?

* Cloudhub is the easiest option for deployment compared to other options
* Each applications runs on separate worker provides application isolation
* Applications can be deployed without any configuration to a hosting environment
* Applications running on multiple workers can share the load automatically and balances the incoming traffic with CloudHub load balancer.
* It monitors services and workers, and if there is any hardware failure, then it shifts to a different worker with zero downtime.
* Provides 99.99% uptime
* One click scalability and automatic update

1. Disadvantages of CloudHub deployment model?

When deploying to CloudHub, keep in mind the following limitations:

* + CloudHub blocks outbound SMTP traffic when more than 20 emails are sent in one hour.
  + CloudHub deployment from Flow Designer fails when the external identity is set up.
  + Only cloudhub administrator can move applications between environment
  + In the cloud, we can have only one application per worker, so we can’t deploy more than ten applications with a Mule instance. However, we can segment out available CPU capacity and monitor resource levels once it’s done.
  + Shared resources for multiple applications in the CloudHub environment is not possible
  + Domain project support is not available
  + Wastage of resource – minimum vcore is .1

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

Default load balancer with cloudHub

1. What are the characteristics of a SLB?

SLB url always have xxx.eu.cloudhub.io

SLB has external IP

SLB do not allow vanity url

Do not have custom option for url redirect

It is shared with multiple instances of application you deploy

It has a rate limit to make a request. After the limit is exceeded the it throws 503 error

1. What is an EC2 instance?

Amazon Elastic Compute Cloud (EC2) is a web service that provides secure, resizable compute capacity in a cloud.

Instances in AWS are basically virtual environments. These virtual environments are isolated from the underlying base OS

EC2 Instances are highly scalable, meaning, you can scale up or scale down based on your requirement dynamically. Using EC2 Instances as your cloud computing environment eliminates the need to invest in hardware and software dependencies.

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

AWS(Amazon Web Service) is the world broadly adopted cloud platform.

AWS is a secure cloud services platform, offering compute power, database storage, content delivery and other functionality to help businesses scale and grow

CloudHub as itself and the Mule applications are deployed and executed under an AWS infrastructure.

Specifically, each deployed application is running under an EC2 instance and a Mule runtime within a JVM.

10.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

11. What is a Worker?

Worker is a dedicated instance of Mule that runs our integration application. Applications on Cloudhub are run by one or more instances of Mule called workers.

12. What is an availability zone?

An available zone is a logical data center is a region available for use by any AWS customer.

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

A region is a physical location around the world where we cluster data centers.

An AWS Region is a geographical location with a collection of availability zones mapped to physical data centers in that region. Every region is physically isolated from and independent of every other region in terms of location, power, water supply, etc.

14. 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

15. What is application isolation?

Application isolation is the separation of one program or application stack from the rest of the running processes.

In Cloudhub each worker runs in a separate container, thus provide isolation from every other application.

16. What is vertical scalability?

Vertical scalability means adding more power or compute resources like memory, CPU in existing servers.

17. How do you achieve vertical scalability?

MuleSoft, Vertical scaling can be achieved by increasing worker size in CloudHub.

Scale Out: Adding more power to your existing machines.

Scale In: Reducing the power of your existing machines.

18. What is horizontal scalability?

Horizontal scaling means adding more number of machines into our pool of resources,

19. 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.

20. What is high availability?

High availability(HA) is the ability of a system or system component to be continuously operational for a desirably long length of time. Availability can be measured relative to 100% operational or never failing

HA is the measure of a system’s ability to remain accessible in the event of a system component failure. Generally, HA is implemented by building in multiple levels of fault tolerance and/or load balancing capabilities into a system.

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

CloudHub provides high availability with the clustering feature, which includes load balancing, persistent message queues, and horizontal scaleout.

CloudHub also actively monitors services and workers for problems and corrects issues. For example, in the case of hardware failure, CloudHub auto-migrates the application to a different worker using CloudHub zero downtime updates.

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

Disaster recovery (DR) is the process by which a system is restored to a previous acceptable state, after a natural or man-made disaster.

By deploying application to multiple regions

23. When should customers choose CloudHub deployment model?

When customer don’t have an infrastructure of their own, 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

The Anypoint Virtual Private Cloud (VPC) offering allows you to create a virtual, private, and isolated network segment in the cloud to host your CloudHub workers.

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

1. What is CloudHub VPC?

The Anypoint Virtual Private Cloud (VPC) offering allows you to create a virtual, private, and isolated network segment in the cloud to host your CloudHub workers.

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

1. Why do you need a VPC?

* Provide vanity url
* Out of box dedicated load balancer
* Provide a private virtual cloud

1. Advantages of having a VPC?

Vanity url

Private connection

DLB

Provides ability to prevent some requests (by blocking implemenataion url and directing to proxies)

1. Disadvantages of having a VPC?

Costly

1. How do you configure a VPC?
2. What is Dedication Load Balancer?

CloudHub dedicated load balancers (DLBs) are an optional component of Anypoint Platform that enable you to route external HTTP and HTTPS traffic to multiple Mule applications deployed to CloudHub workers in a Virtual Private Cloud (VPC).

Dedicated load balancers enable you to:

* Handle load balancing among the different CloudHub workers that run your application.
* Define SSL configurations to provide custom certificates and optionally enforce two-way SSL client authentication.
* Configure proxy rules that map your applications to custom domains.
* It has internal and external IP
* Used for url redirect

This enables you to host your applications under a single domain.

1. What are the features or characteristics of DLB?

CloudHub dedicated load balancers (DLBs) are an optional component of Anypoint Platform that enable you to route external HTTP and HTTPS traffic to multiple Mule applications deployed to CloudHub workers in a Virtual Private Cloud (VPC).

Dedicated load balancers enable you to:

* Handle load balancing among the different CloudHub workers that run your application.
* Define SSL configurations to provide custom certificates and optionally enforce two-way SSL client authentication.
* Configure proxy rules that map your applications to custom domains.

This enables you to host your applications under a single domain.

1. Differences between SLB and DLB?

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. SLB supports HTTPS protocol but the user can not add it's custom certificates, sub-domains and can't expose custom vanity domain names.

A dedicated load balancer is not shared resource in CloudHub and resides inside the client's VPC, user can configure more than one DLB in VPC. As a part of a client's VPC users can configure and customize it. Users can assign n number of workers for high availability, upload custom certificates, upload client certificate, configure custom vanity domain, IP whitelist/blacklist and configure sub-domains. As DLB is not a shared resource in CloudHub, So users can write custom URL mapping rules and run multiple API endpoints under the same domain name.

1. Advantages of having a DLB?

Dedicated load balancers enable you to:

* Handle load balancing among the different CloudHub workers that run your application.
* Define SSL configurations to provide custom certificates and optionally enforce two-way SSL client authentication.
* Configure proxy rules that map your applications to custom domains.
* Used for IP white listing
* It has internal and external IP

This enables you to host your applications under a single domain.

1. What are the ways that we can create a secure connection between CloudHub worker and customer’s networks?

You can connect on-premises data centers through a secured VPN tunnel, or a private AWS VPC through VPC peering, or by using AWS Direct Connect.

1. What is an IPsec tunnel?

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

1. What is direct connect?
   * AWS Direct Connect lets you establish a dedicated network connection between your network and one of the AWS Direct Connect locations.
   * It uses dedicated, private network connections between your intranet and Amazon VPC.
2. What is VPC peering?

A VPC Peering is a network connection between two VPCs that allows traffic routing between the VPCs using private IPv4 or IPv6 addresses.

1. What is a vanity URL?

With dedicated load balancer we can create customized url in VPC deployment model. This customized url (url for branding) is called vanity url

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

With the help of dedicated load balancer

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

When customer needs a direct connection with pvt data centres, customer needs vanity url,

Security purposes, We can maintain on our own..so flexibility is provided

1. What are the firewall rules of VPC?

When you create an Anypoint Virtual Private Cloud (Anypoint VPC), you can add your own firewall rules to allow specific IP ranges and ports to reach your workers.

By default, all traffic to your VPC is blocked unless it’s explicitly allowed in a firewall rule. When you create an Anypoint VPC, four firewall rules are created by default:

* Two rules to allow inbound connections from within your local Anypoint VPC through ports 8091 and 8092
* Two rules to allow inbound connections from anywhere through ports 8081 and 8082

**Hybrid deployment**

1. What is the hybrid type of deployment model?

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?

In hybrid model application can be deployed to

Server

Server group

Cluster

1. Advantages of hybrid deployment model?

The hybrid integration model also provides a [secure data gateway](https://www.mulesoft.com/platform/soa/mule-enterprise-security)which allows for protected communication between CloudHub and Mule as an ESB.

High Availibility.

Usage of domain project.

No core limitation

Reduces licensing cost

1. Disadvantages of hybrid deployment model?

* Load balancing is not provided for hybrid deployments. You can manage load balancing with the tools connected to your on-premises resources.
* The Runtime Manager logging feature is not available for hybrid deployments. You can configure on-premises apps to send data to external analytics tools, such as Splunk or ELK, to manage your logs.
* The object store infrastructure is not available for hybrid deployments. To use object stores, you must configure a database to store data.
* After an application is deployed and running, you must apply any security updates for the selected runtime version manually.

1. What is a standalone deployment?

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

1. What is a cluster?

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. In simple terms, virtual servers composed of multiple nodes and they communicate and share information through a distributed shared memory grid.

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. If you experience concurrency issues with server groups, consider grouping servers in clusters.

1. What is a hazelcast?

In a Hazelcast grid, data is evenly distributed among the nodes of a [computer cluster](https://en.wikipedia.org/wiki/Computer_cluster), allowing for [horizontal scaling](https://en.wikipedia.org/wiki/Horizontal_scaling) of [processing](https://en.wikipedia.org/wiki/Parallel_computing) and available storage. Backups are also distributed among nodes to protect against failure of any single node. Hazelcast provides central, predictable scaling of applications through in-memory access to frequently used data and across an elastically scalable data grid. These techniques reduce the query load on databases and improve speed.

1. How does node replication happen in a hybrid-based deployment model?
2. What is the use of node replication?

For HA, it reduces the workload on a single node

1. What distributed memory in a cluster-based deployment and how it is useful?

In a cluster based environment, resources and work load are shared between nodes and replicated for high availability. This is achieved internally by hazelcast implemented by mulesoft.

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

A unicast cluster requires that we configure the IP addresses of the nodes in the cluster. If a server has multiple interfaces, use the internal IP address that allows the node to communicate directly with other nodes. Clustering across different subnets is not supported.

Because Mule relies on the IP address to identify the server, IP addresses can’t be dynamically assigned using DHCP for servers on a unicast cluster. If a server is restarted and uses DHCP to get a new IP address, we must add the server to the cluster using its new IP address.

The server status must be Running to be added to a unicast 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 we can add nodes to the cluster dynamically without restarting the cluster.

It is necessary to check with the network administrator to determine if multicast is allowed within our network, because many networks block multicast functionalities.

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

In cluster based deployment application will be deployed to different servers in the same cluster, so if one of the server node is down high availability can be achieved as the application will be still running from other server node

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

Load balancing is not provided for hybrid deployments. We can manage load balancing with the tools connected to our on-premises resources like nginx, HAProxy.

1. How can we achieve high availability in hybrid?

To achieve high availability in hybrid deployments, use server groups or clusters. Clustered Mule instances have distributed shared memory, which provides persistent VM queues, transactions, and cluster-wide data storage.

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

Horizontal scalability can be achieved by increasing number of nodes (servers)

Vertical scalability can be achieved by increasing memory size and processor speed

**Runtime Fabric (RTF)**

1. What is runtime fabric type of deployment?

RTF is a deployment model in which application will run in a customer hosted runtime plane and it can be managed by the mulesoft managed control plane

Runtime Fabric (RTF) is a container service to deploy mule runtime on cloud (AWS / Azure) or on a data center (on-premise)

1. When should a customer choose RTF deployment model?

A customer chooses RTF deployment model when want to have high availability, scalability, application isolation features of cloudhub and application to be deployed in their own private cloud that could be in AWS, Azure or on-premise

1. Advantages of using RTF?

* Multicloud: RTF can be installed to any cloud or customer datacentre and managed via the Anypoint Platform Runtime Manager
* Application Isolation: With underlying Docker container technology, it provides application isolation
* Mutiple Runtimes: Because of Isolation it is possible to deploy multiple version of mule runtime to same cluster. Customers can run Mule 3 and Mule 4 application in one infrastructure
* Zero Downtime Upgrades: Customers can enjoy painless automation like cloudhub
* High availability: RTF cluster installations are highly available by design, and deploying Mule applications in a highly available replica set or cluster is simple as moving a slider and clicking a checkbox.
* Minimal Infrastructure Skills: underlying docker and kubernetes technologies are mostly transparent to end user

1. Disadvantages of using RTF?

* RTF only works with MuleSoft-hosted AnyPoint Platform (US or EU). There is currently no support for customer-hosted control-planes.
* There is no native support for ObjectStore or local storage. An external DB or file server should be used for persistent storage of application data.
* No support for domain project.

1. Differences between RTF and CloudHub deployment model?

In cloudhub kind of deployment everything is managed by mule, i.e both control plane as well as runtime plane is managed by mule, but in RTF runtime is managed by customer and cloudhub features are can be achieved here easily.

1. Differences between RTF and Hybrid deployment model?

In hybrid deployment updations have to be made manually, we need to setup load balancer manually, we can’t deploy mule 3 and mule 4 apps within the same resources, application share same resource so all app are prone to downtime.

1. What is a Docker container?

Docker is a tool designed to make it easier to create, deploy, and run applications by using containers.

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 a container based technology and Virtual machine is not based on container technology

In Docker, Containers running share the host OS kernel. Each VM has OS and apps, it shares hardware resource from the host

We can use Docker to isolate individual applications and use Virtual Machines to isolate entire systems

1. What is Kubernetes?

It's is an open-source container orchestration tool designed to automate deploying, scaling, and operating containerized applications.

1. Features of Kubernetes?

* Automated Scheduling
* Self-Healing Capabilities
* Automated rollouts & rollback
* Horizontal Scaling & Load Balancing
* Offers environment consistency for development, testing, and production
* Infrastructure is loosely coupled to each component can act as a separate unit
* Provides a higher density of resource utilization
* Offers enterprise-ready features
* Application-centric management
* Auto-scalable infrastructure
* You can create predictable infrastructure

1. What is a controller?

Controller node is the Kubernetes controller which will do internal load balancing and most control plane communication with Anypoint Platform happens here

The maximum number of supported controller node is 5

1. What are the characteristics of a controller?

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.

1. What is a worker node?

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.

The current maximum number of supported worker nodes is 16.

1. What is a POD?

Pods are the smallest deployable units of computing that you can create and manage in Kubernetes. A Pod (as in a pod of whales or pea pod) is a group of one or more containers, with shared storage/network resources, and a specification for how to run the containers

1. How do you achieve application isolation in RTF?

In RFT applications are deployed to pods in worker node. Thus each application can be isolated from each other

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

Horizontal scalability is achieved by making replicas of the deploying app.

Vertical scalability can be achieved here by increasing the vCore size.

1. How do you achieve high availability in RTF?

Runtime Fabric provides automatic application fail-over. Also, Runtime Fabric provides high availability when you deploy two or more replicas per application.

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

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?

In production 3 controller nodes and 3 worker nodes

In Development 1 controller node and 2 worker node

1. How do you balance the load in RTF?

Runtime Fabric includes an internal load balancer for basic load balancing. You must connect an external load balancer for production configurations.

1. What control plane capabilities are available for RTF?

Private Cloud Edition (PCE)

1. What is PCE?

PCE is a deployment model in which both control plane and runtime plane are managed by customer.

Using PCE one can run and manage Mule applications on their local servers using their required security policies.

It enables to maintain control over data storage and processing of data

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

If an organization has a strict compliance law on data processing within their own data centers then PCE is suitable. In such a case the organizations won’t even allow the flow of meta data outside the premise. Example industries such as banks, insurance corps, healthe care and government organization have scenarios where they want to more tightly manage their data.

1. Advantages of PCE?

* Control Plane and Runtime plane are secured by the organisation
* No flow of meta data to the Mulesoft hosted cloud
* All the exchange assets are strictly remain on premise

4. Disadvantages of PCE?

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

1. What are the differences between RTF and PCE?

* 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?

Available components are Anypoint Runtime Manager, Anypoint API Manager, Anypoint Exchange, Access Management, Anypoint Design Centre excluding Flow Designer,

1. What are the basic configurations of a PCE?

3 Node Configuration

Three is the minimum number of nodes required to provide high availability and application fail over

Each node will have database instance and object store of its own and But there will be only one master database

Load balancer should be configured

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?

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 runtime plane is the collection of the clusters.

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

In this the control plane components like design center and management are released into docker containers. Those docker containers are managed by the Kubernetes.

Control plane uses Docker and Kubernetes

1. Does PCE support application isolation and why?

Yes, PCE supports application isolation because applications are deployed to Docker containers

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

Horizontal scaling can be achieved by increasing the number of pods/servers

Vertical scaling can be increased by increasing the resources

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. In case of one server failure still the request from the user can be completed by another backup server.  Thus it achieves zero down time

1. How do you balance the load in PCE?

PCE run in a production environment with multiple servers. To distribute traffic among servers and to restrict access to specific ports, we need to install and configure a load balancer

You can configure your load balancer to use any method for distributing client requests, but in most contexts a round robin strategy is ideal. This load balancer should be reachable through an IP address that is accessible to all machines in your network.

Your load balancer must route the following TCP ports:

In each case, your load balancer must listen on the Load Balancer Port and redirect incoming requests to the Instance Port. Your Anypoint Platform installation includes an internal NGINX server that listens on each of the configured instance ports, then performs the action listed in the Internal Usage column.

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

  PCF is a cloud native platform for deploying and operating modern applications.

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

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

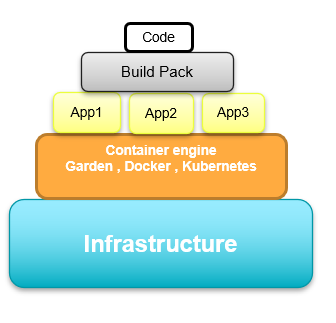
1. Advantages of PCF?

* Latency: when our core systems are in one data centre and we want our runtime to be in same data centre
* Compliance requirements: Government specific projects
* GDPR and Data residency requirement

1. Disadvantages of PCF?

* Customer need to have extra configuration for
  + High availability
  + Logging: need to leverage third party logging service like Splunk
  + Load balancer
  + Monitoring: we can only trigger Alerts, not able to see utilization information like in Cloud hub control plane
* Need to manually update security updates

1. Explain the basic architecture of PCF.



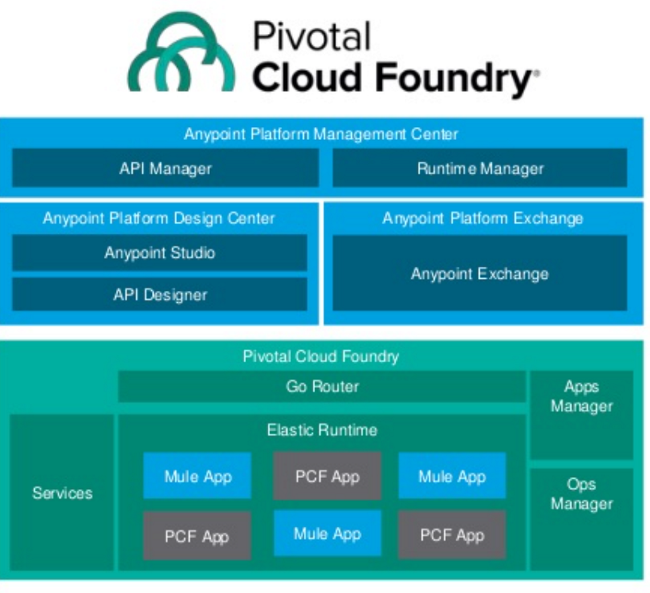
* PCF can host both control plane and Runtime plane
* PCF can be hosted on customer Data center or AWS or GCP or on any cloud providers
* If we have an infrastuctutre On top of ‘Infrastructure’ ,PCF provides own container engine which is like Garden or Docker or Kubernetes
* On top of container engine, we deploy our containerized applications.
* PCF determines mule applications with the help of ‘Build Pack’
* What usually happens is whenever you create a code, the code has something 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 centre
* Unlike manual runtime installation in PCE, runtimes are provisioned automatically using PCF technology.
* Unlike PCE, every application will get separate runtime in PCF

Buildpacks:

* 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.

**Integration Architecture:**



  Anypoint Platform for PCF uses the following services:

**Runtime Manager Service**: exposes PCF as a deployment target within the Anypoint Runtime Manager and lets you deploy Mule applications to PCF, alongside other deployment targets.

**API Gateway Service**: enables you to create and deploy a Mule API proxy that intercepts traffic to a non-Mule application running on PCF.

MuleSoft interface to apply policies over the API and the PCF interface to view usage and policy metrics.

**Anypoint Metering Service**: provides an API that enables you to view usage metrics for Mule applications deployed to PCF.

You can view usage metrics for production, pre-production, staging, and development environments.

1. What are the differences between RTF and PCF?

In PCF customer manage both control plane and runtime plane

1. What are the differences between PCE and PCF?

Unlike manual runtime installation in PCE, runtimes are provisioned automatically using PCF build pack technology.

Unlike PCE, every application will get separate runtime in PCF , thus we will get application isolation

1. What control plane capabilities are available for PCF?

* Runtime Manager Service:
  + exposes PCF as a deployment target
  + let us deploy Mule application
* Application gateway
  + Enable to create and deploy Mule proxy applications
* Application metering service
  + Enables to view usage metrics for Mule applications

1. What are droplets in PCF and how is it used?

* Build pack will 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.
* 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 by deploying application to droplets

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?
2. How do you achieve zero-downtime in PCF?

Customer need to configure

1. How do you balance the load in PCF?

Customer need to configure load balancer

**Runtime Behaviour / Thread pool**

1. What is a thread?

A thread is a unit of execution. It is the virtual version of a CPU core

2. What is a thread pool?

A thread pool is a collection of available threads

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

SEDA (Staged Event Driven Architecture) 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 change load

4. What is reactive programming?

**Reactive programming**is a programming paradigm that deals with asynchronous data streams (sequences of events) and the specific propagation of change, which means it implements modifications to the execution environment (context) in a certain order.

5. How reactive programming is used in mule?

Mule 4 uses reactive programming to facilitate non-blocking execution of the event processors. This has a significant impact on a Mule application’s ability to scale the amount of events it can handle concurrently.

6. What are the features of reactive programming?

Features:

* 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.

7. 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 signal to the producer to “please slow down.” This process is known as back pressure

8. How backpressure is implemented in mule?

Backpressure is implemented either automatic or manual way in mule 4

Automatic backpressure:

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.

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 pass through the event processor per second.

9. 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.

10. 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 pass through the event processor per second

11. How to configure a manual backpressure?

Developers can configure maxConcurrency attribute to achieve manual backpressure

12. 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.

13. What are the different processing types in Mule?

Event processors in Mule is categorized into 3 different Processing types

CPU\_LITE

* For tasks that take up to 10ms to execute.
* No blocking IO operations should be executed here
* Default size is 2 \* cores
* Logger, HTTP Requester are examples

CPU\_INTENSIVE

* + - * 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.
      * Non-blocking IO operations should be executed here.
      * Default size is 2 \* cores.
      * Dataweave, Scripting are examples

IO\_BLOCKING

* 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, IO are examples

14. What is CPU\_LITE processing type?

Event processors that take less than 10ms to complete processing are called CPU\_LITE Processing type

15. Examples of connectors/processors which use CPU\_LITE processing type?

Logger

Set payload

Http Request

16. What is CPU\_INTENSIVE processing type?

Event processors which perform non-blocking operations and which take more than 10ms to complete. It requires more memory than CPU\_LITE

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

Transform Component

Scripting like groovy

18. What is the BLOCKING\_IO processing type?

Event processors that perform blocking operations.

Significantly larger than the other pools, as most threads here are expected to be in a blocked state.

19. Examples of connectors/processors which use BLOCKING\_IO processing type?

Database

IO

20. 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

21. 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

22. 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

23. 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

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

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).

25. 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 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

26. What is thread switching?

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

27. 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.

28. 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.

29. 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

30. 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

31. How does mule increase t

32. 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

33. 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

34. 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