128

Self Healing Applications

Liveness probe, Readiness probe not part of CKA.

129. Introduction to Autoscaling

* Horizontal Pod Autoscaling (HPA)
* Vertical Pod Autoscaling (VPA)

Vertical Scaling

Back in time, apps hosted on physical servers with preconfigured CPU, memory capacity, & what happens when load increases & we run out of existing resources on server?

we need to scale up the server, so we took down the application & added more CPU (or) memory resources to it & then powered it back up. This is Vertical Scaling

⇒ Basically increase the size of the existing server Vertically we added more resources to our existing server by adding more CPU & memory component to it Vertically

Horizontal Scaling

If the application supported running on multiple instances another thing we could have avoided having to shutdown the server, Instead we could have added more servers to it. & shared load b/w them

Running more instances of apps by adding more servers is known as horizontal scaling

Vertical Scaling ⇒ Adding more resources to existing application

Horizontal Scaling ⇒ Adding more instances or more servers to your system

2:10

One of the major purpose of a container orchestrator like K8S is to host apps in the form of containers & scale up and down based on demands

Scaling
1- Scaling Cluster Infra — VM
2- Scaling Workloads. — K8S

Scaling workloads

We can scale the workloads by adding or removing containers or pods on to the cluster.

Scaling underlying cluster { Scaling cluster Infra }

adding or removing more servers or infrastructure to your cluster

2:59

Scaling Cluster Infra :-

Types
1) Horizontal Scaling 2) Vertical Scaling

Horizontal Scaling cluster :-

refers to adding more nodes to the cluster

Vertical Scaling cluster :-

Increasing the resources on existing nodes in the cluster

Scaling work loads

* Horizontal Scaling * Vertical Scaling

Horizontal Scaling :-

Creating more pods

Vertical Scaling :-

Increasing the resources allocated to existing pods

How to Scale

There are 2 ways of Scaling

* Manual way * Automated way.

Manual way of horizontally Scaling cluster infra

⇒ Manually provision new nodes

⇒ Then use the kubeadm join ... command to add new nodes to the cluster

when its come to **Vertical Scaling** there is no Common approach used for k8s because it will result in having to take down the servers & apps running on them & then add more resources to the servers & bring it backup. This we dont want to do. Since most of the infra these days are VMs, we could easily provision a server with higher resources, add it to the cluster & remove the old server. So Vertical scaling is not common approach

Horizontal Scaling workloads manual approach

Run the kubectl scale command on the workload to scale up or down the no of pods

Vertical Scaling workload manual approach

Vertically scaling resources associated to a pod would usually run the kubectl edit command to go into that deployment or statefulset or replicaset & change the resource limits & requests associated with the pods

~~Automated approach for horizontally Scaling the infra~~

Automated approach:-

⇒ For horizontally Scaling the infra, we have what is known as the K8S cluster autoscaler.

⇒ we can't do anything in automated way in terms of vertical scaling in infra as we have to manually shutdown, scale up & power it up.

⇒ For horizontally Scaling the workloads, we have what is known as the **Horizontal Pod Autoscaler or HPA**

⇒ For Vertically Scaling the workloads, we have the **Vertical Pod Autoscaler**