**Infra Optimization.**

Project 1 **Description**

**Background of the problem statement:**

A popular payment application, EasyPay where users add money to their wallet accounts, faces an issue in its payment success rate. The timeout that occurs with the connectivity of the database has been the reason for the issue. While troubleshooting, it is found that the database server has several downtime instances at irregular intervals. This situation compels the company to create their own infrastructure that runs in high-availability mode. Given that online shopping experiences continue to evolve as per customer expectations, the developers are driven to make their app more reliable, fast, and secure for improving the performance of the current system.

**Solution**

Create a DevOps infrastructure for an e-commerce application to run on **HA** (high availability) node.

**High-Level Details**

The application will run on Amazon’s Elastic Kubernetes Service (EKS). EKS is a robust commercial solution that is reliable and easily scales. Key to the service is the deployment of a database, running inside a Pod, that can easily be restarted for any reason. The data in the database must be persistent, i.e., not lost if the server goes down. So, in this solution the database pod uses a volume tied to the hardware of a particular node. Since we separate the backend hardware solution from the front-end, the vendor who employs this solution can decide how much to invest in network, storage, and processing for the database server.

**Details**

In order to get this working, we need to install the AWS CLI, EKS, and Ansible on a Mac, PC, or Linux system. (In my particular case I setup a Debian VM running in VirtualBox on my Macbook Pro.) With these tools in place we are ready to launch the Kubernetes cluster on AWS.

EKS requires an AWS account, which has been previously setup. Next we launch the EKS cluster using the EKS CLI, eksctl, as follows:

eksctl create cluster \

--name evansjr \

--region us-west-2 \

--nodes 3 \

--nodes-max 5 \

--with-oidc \

--ssh-access \

--managed

(On average it takes Amazon about twenty minutes to setup the cluster.)

Once the Kubernetes cluster is up and running, the nodes are provisioned with Ansible. This requires that we be able to SSH into the nodes, which EKS allows using PKI certificates. All the ansible scripts are available at the github repository for this Project: **http://github.com/evansjr2000/Capstone.git**, and some examples are given in the appendix.

**Setting up a Persistent Volume**

The database will reside on a specific node of the cluster. To prevent the database pod running on a node not containing the database storage we will label the node.

kubectllabel node$(kubectl getnodes -ojsonpath=’{.items[0].metadata.name}’) kiamol=ch05

Here is the specification for the persistent volume, which will attach to the node with the correct label, as defined above.

apiVersion: v1 : PersistentVolume metadata:

name: pv01 spec:

capacity:

storage: 50Gi

accessModes:

* ReadWriteOnce local:

path: /nfs\_share\_dir nodeAffinity:

required:

nodeSelectorTerms:

* matchExpressions:
* key: kiamol operator: In values:
* ch05

And here is the manifest for the Persistent Volume Claim.

apiVersion: v1 kind: PersistentVolumeClaim metadata: name: postgres-pvc spec:

accessModes:

* ReadWriteOnce resources:

requests:

storage: 40Mi storageClassName: ""

The access mode of **ReadWriteOnce** limits access to this volume to one pod, in this case the database pod. As previously mentioned the database pod will run on a specified node of the cluster.

Here are the manifests for the database service and deployment;

apiVersion: v1 kind: Service metadata:

name: todo-db spec:

ports:

- port: 5432 targetPort: 5432

selector:

app: todo-db

type: ClusterIP

--apiVersion: apps/v1 kind: Deployment metadata:

name: todo-db spec:

selector:

matchLabels:

app: todo-db template:

metadata:

labels: app: todo-db spec:

containers:

* name: db image: postgres:11.6-alpine env:
* name: POSTGRES\_PASSWORD\_FILE value: /secrets/postgres\_password volumeMounts:
* name: secret mountPath: "/secrets"
* name: data mountPath: /var/lib/postgresql/data

volumes:

* name: secret secret:

secretName: todo-db-secret defaultMode: 0400 items:

* key: POSTGRES\_PASSWORD path: postgres\_password
* name: data persistentVolumeClaim: claimName: postgres-pvc

To access the database we use a web front-end. This particular example uses a LoadBalancer manifest and we allow it to scale, using a Horizontal Pod Autoscaler (HPA) manifest as follows:

apiVersion: autoscaling/v1 kind: HorizontalPodAutoscaler metadata: name: todo-web labels: kiamol: ch19 spec:

scaleTargetRef: apiVersion: apps/v1 kind: Deployment name: todo-web minReplicas: 1 maxReplicas: 5

targetCPUUtilizationPercentage: 10

The target CPUUtilizationPercentage is set artificially low above in order to demonstrate autoscaling when the from end web server undergoes a large load.

The following is the manifest for the web service:

apiVersion: v1 kind: Service metadata: name: todo-web spec:

ports:

- port: 8081 targetPort: 80

selector:

app: todo-web

type: LoadBalancer

The manifest for the deployment is very large and can be found in the appendix.

**Testing**

1. To verify that the database robustly restarts after dying (or being deleted) and that the backend data is preserved (persistent) we show what is in the database before deleting the service, then as the service will automatically start another deployment of the database, we show that the data is still there. Xxxxx

**Implementation requirements:**

1. Create the cluster (EC2 instances with load balancer andelastic IP in case of AWS)

Kubernetes provides the load balancer. An Elastic IP address is a static IPv4 address designed for dynamic cloud computing. An Elastic IP address is allocated to your AWS account, and is yours until you release it. Or maybe not.

(See https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/elasticip-addresses-eip.html)

1. Automate the provisioning of an EC2 instance using Ansibleor Chef Puppet

(I have scripts to do this.)

1. Install Docker and Kubernetes on the cluster (I have scriptsto do this.)
2. Implement the network policies at the database pod to allowingress traffic from the front-end application pod (Need to look this up in KIAMOL book.)
3. Create a new user with permissions to create, list, get, update, and delete pods (Doable. Must research and demonstrate how to do this.)
4. Configure application on the pod (Maybe implement thedatabase.)
5. Take snapshot of ETCD database (ETCD is the key-valuestore for K8s. Need to read about “backing up K8s.”
6. Set criteria such that if the memory of CPU goes beyond

50(Did this as part of last class. Need to re-implement this.)

**The following tools must be used:**

1. EC2
2. Kubernetes
3. Docker
4. Ansible or Chef or Puppet

**The following things to be kept in check:**

1. You need to document the steps and write the algorithms inthem.
2. The submission of your GitHub repository link is mandatory.In order to track your tasks, you need to share the link of the repository.
3. Document the step-by-step process starting from creatingtest cases, then executing them, and recording the results.
4. You need to submit the final specification document, whichincludes:
   1. Project and tester details
   2. Concepts used in the project
   3. Links to the GitHub repository to verify the projectcompletion
   4. Your conclusion on enhancing the application and defining the USPs (Unique Selling Points)

**Appendix**

$ kubectl get nodes -o wide

NAME STATUS ROLES AGE

VERSION INTERNAL-IP EXTERNAL-IP OS-IMAGE KERNEL-VERSION

CONTAINER-RUNTIME ip-192-168-1-152.us-west-2.compute.internal Ready <none> 2d2h v1.19.6-eks-49a6c0 192.168.1.152 34.216.36.123 Amazon Linux 2 5.4.117-58.216.amzn2.x86\_64 docker://19.3.13

ip-192-168-57-176.us-west-2.compute.internal Ready <none> 2d2h v1.19.6-eks-49a6c0 192.168.57.176 35.80.10.134 Amazon Linux 2 5.4.117-58.216.amzn2.x86\_64 docker://19.3.13

ip-192-168-95-52.us-west-2.compute.internal Ready <none> 2d2h v1.19.6-eks-49a6c0 192.168.95.52 34.221.188.50 Amazon Linux 2 5.4.117-58.216.amzn2.x86\_64 docker://19.3.13

Using the above information I create a hosts file for Ansible to access the Kubernetes nodes:

$ cat hosts [k8s-nodes] k8smaster01 ansible\_host=34.216.36.123 ansible\_user=ec2-user k8snode01 ansible\_host=35.80.10.134 ansible\_user=ec2-user k8snode02 ansible\_host=34.221.188.50 ansible\_user=ec2-user

[all:vars]

# Debian uses python3 instead of python2, # we need to set the interpreter for ansible ansible\_python\_interpreter=/usr/bin/python

[kube-master] k8smaster01

[kube-node] k8snode01 k8snode02

Here is the manifest for the deployment of the web service:

apiVersion: apps/v1 kind: Deployment metadata: name: todo-web spec:

selector: matchLabels: app: todo-web template:

metadata:

labels:

app: todo-web spec:

containers:

* name: web image: kiamol/ch04-todo-list resources:

limits:

cpu: 250m

requests:

cpu: 125m

env:

* name: ASPNETCORE\_ENVIRONMENT value: Test volumeMounts:
* name: config mountPath: "/app/config" readOnly: true
* name: secret mountPath: "/app/secrets" readOnly: true

volumes:

* name: config configMap: name: todo-web-config items:
* key: config.json path: config.json
* name: secret secret:

secretName: todo-web-secret defaultMode: 0400 items:

* key: secrets.json path: secrets.json