9. Autoscaling

Включить метрики в API-server

> via https://kubernetes.io/docs/tasks/run-application/horizontal-pod-autoscale/

1. Добавить флаг **--enable-aggregator-routing=true** в файл **/etc/kubernetes/manifests/kube-apiserver.yaml** на всех **control plane** нодах
2. Зарегистрировать API для метрик ресурсов, подняв сервер метрик:

   ```
   git clone https://github.com/kubernetes-incubator/metrics-server.git
   ```

   добавить следующий блок в файл metrics-server/deploy/1.8+/ в описание контейнера:

   ```
   command:
             - /metrics-server
             - --logtostderr
             - --v=3
             - --metric-resolution=30s
             - --kubelet-insecure-tls
             -
   --kubelet-preferred-address-types=InternalIP,ExternalIP,Hostname
   ```

   и применить описания:

   ```
   kubectl apply -f metrics-server/deploy/1.8+/
   ```

Изменить задержки autoscale/downscale для Controller manager

Добавить в файл **/etc/kubernetes/manifests/kube-controller-manager.yaml** в параметры запуска две строки:

```
- --horizontal-pod-autoscaler-downscale-delay=2m30s
    - --horizontal-pod-autoscaler-upscale-delay=0m30s
```

Добавить HPA

HPA = Horizontal Pod Autoscaler

> via https://kubernetes.io/docs/reference/generated/kubectl/kubectl-commands#autoscale

https://kubernetes.io/docs/tasks/run-application/horizontal-pod-autoscale-walkthrough/

```
$ kubectl autoscale deployment work-order-parser --min=2 --max=10
--cpu-percent=90
horizontalpodautoscaler.autoscaling/work-order-parser autoscaled
$ kubectl get hpa
NAME                REFERENCE                     TARGETS   MINPODS
MAXPODS   REPLICAS   AGE
work-order-parser   Deployment/work-order-parser   0%/90%    2         10
2         22m
$ kubectl describe hpa work-order-parser
Name:                                                      work-order-parser
Namespace:                                                 robotization
Labels:                                                    <none>
Annotations:                                               <none>
CreationTimestamp:                                         Wed, 19 Sep 2018
10:54:20 +0300
Reference:
Deployment/work-order-parser
Metrics:                                                   ( current / target )
  resource cpu on pods  (as a percentage of request):  0% (2m) / 90%
Min replicas:                                              2
Max replicas:                                              10
Deployment pods:                                           2 current / 2
desired
Conditions:
  Type            Status  Reason            Message
  ----            ------  ------            -------
  AbleToScale     True    ReadyForNewScale  the last scale time was
sufficiently old as to warrant a new scale
  ScalingActive   True    ValidMetricFound  the HPA was able to
successfully calculate a replica count from cpu resource utilization
(percentage of request)
  ScalingLimited  True    TooFewReplicas    the desired replica count is
increasing faster than the maximum scale rate
Events:
  Type     Reason                        Age                 From
Message
  ----     ------                        ----                ----
-------
  Warning  FailedComputeMetricsReplicas  16m (x13 over 22m)
horizontal-pod-autoscaler  failed to get cpu utilization: unable to get
metrics for resource cpu: unable to fetch metrics from resource metrics
API: the server is currently unable to handle the request (get
pods.metrics.k8s.io)
  Warning  FailedGetResourceMetric       12m (x21 over 22m)
horizontal-pod-autoscaler  unable to get metrics for resource cpu: unable
to fetch metrics from resource metrics API: the server is currently unable
to handle the request (get pods.metrics.k8s.io)
```