Spark Operator

1. Path to Spark Configurations

   The configuration files for Spark are located at: /root/cafebot-kube/spark-operator.

2. Building Custom Spark Image

   To build a custom Spark image with Spark app and custom JAR libraries, use the following Dockerfile. Ensure the Spark app and required JARs are placed in /root/cafebot-kube/spark-operator/app and /root/cafebot-kube/spark-operator/jars respectively before building and pushing the image:

   Dockerfile

```
#Dockerfile
# Use the official Apache Spark PySpark image as the base
FROM apache/spark-py:latest

# Copy the application files to the work directory
COPY app /opt/spark/work-dir/
COPY jars /opt/spark/jars

# Switch to a non-root user
USER 185:0

# Switch to root to download Microsoft JDBC Driver for SQL
USER root
#RUN wget -O /opt/spark/jars/mssql-jdbc-9.2.1.jre8.jar htt

# Switch back to the non-root user



# Install Python dependencies with --user flag
#RUN pip install --no-cache-dir -r /opt/spark/work-dir/req
```

```
# Change ownership of the installed packages directory
RUN chown -R 185:0 /opt/spark/work-dir
RUN chown -R 185:0 /opt/spark/jars
USER 185:0
```

build and push the image

```
docker build . -t repo:tag
docker push repo:tag
```

> 💡 **In this case, the tag is shekharzxcv/spark-py:latest**

3. Spark Application

```
apiVersion: "sparkoperator.k8s.io/v1beta2"
kind: SparkApplication
metadata:
  name: pyspark-pi
  namespace: airflow
spec:
  type: Python
  pythonVersion: "3"
  mode: cluster
  image: "shekharzxcv/spark-py:remote"
  imagePullPolicy: Always
  mainApplicationFile: local:///opt/spark/work-dir/main.py
  sparkVersion: "3.5.0"
  sparkConf:
    "spark.eventLog.enabled": "false"
    "spark.eventLog.dir": "file:/mnt"
  restartPolicy:
    type: OnFailure
    onFailureRetries: 3
    onFailureRetryInterval: 10
    onSubmissionFailureRetries: 5
    onSubmissionFailureRetryInterval: 20
  volumes:
    - name: spark-data
      persistentVolumeClaim:
        claimName: spark-pvc
  driver:
    cores: 1
    coreLimit: "1200m"
    memory: "512m"
      #hostNetwork: true
    labels:
      version: 3.1.1
    serviceAccount: spark-spark-operator
    volumeMounts:
      - name: spark-data
        mountPath: /mnt
  executor:
    cores: 1
    instances: 3
    memory: "512m"
    volumeMounts:
      - name: spark-data
        mountPath: /mnt

  sparkUIOptions:
    ingressAnnotations:
      kubernetes.io/ingress.class: nginx
  dynamicAllocation:
    enabled: true
    minExecutors: 1
    maxExecutors: 10
```

4. Scheduled Spark Application

```
apiVersion: "sparkoperator.k8s.io/v1beta2"
kind: ScheduledSparkApplication
metadata:
  name: spark-pi-scheduled
  namespace: cafebot2
```

```
spec:
  schedule: "@every 10m"
  concurrencyPolicy: Allow
  successfulRunHistoryLimit: 1
  failedRunHistoryLimit: 3
  template:
    type: Python
    mode: cluster
    image: "shekharzxcv/spark-py:latest"
    #mainClass: org.apache.spark.examples.SparkPi
    sparkVersion: "3.5.0"
    mainApplicationFile: local:///opt/spark/work-dir/main.
    driver:
      cores: 1
      coreLimit: "1200m"
      memory: "512m"
      #hostNetwork: true
      labels:
        version: 3.1.1
      serviceAccount: spark-spark-operator
      volumeMounts:
        - name: spark-data
          mountPath: /mnt
    executor:
      cores: 1
      instances: 1
      memory: "512m"
      volumeMounts:
        - name: spark-data
          mountPath: /mnt
    restartPolicy:
      type: OnFailure
      onFailureRetries: 3
      onFailureRetryInterval: 10
      onSubmissionFailureRetries: 5
      onSubmissionFailureRetryInterval: 20
    volumes:
      - name: spark-data
```

```
        persistentVolumeClaim:
          claimName: spark-pvc

    sparkUIOptions:
      ingressAnnotations:
          kubernetes.io/ingress.class: nginx
    dynamicAllocation:
      enabled: true
      minExecutors: 3
      maxExecutors: 10
```

5. Spark helm chart deployment

   To deploy Spark using the Helm chart, navigate to /root/cafebot-kube/spark-operator and run the following commands:

   Initial Deployment

   ```
   helm install spark . -n cafebot2 -f values.yml
   ```

6. Upgrading Helm Chart

   ```
   helm upgrade spark . -n cafebot2 -f values.yml
   ```