Deploying HPC platforms using ProActive

The main steps

1. **Swarm.xml** to allow container-to-container network using overlay network, natively supported by docker engine swarm mode (a swarm is a cluster of Docker engines/nodes offering services: add/remove nodes,..). An overlay network requires a key-value store (here consul supported by docker) to holds information about the network state.

1. Deploy your HPC platform via docker containers, for a portable deployment. To start/control docker containers on (remote) hosts, we use the docker-machine command relying on ssh.

   a. **HDFS.xml** if you just need a dedicated file system adapted to big data
   b. **Spark.xml** for a big data processing platform
   c. **HDFS.xml** and **Spark.xml** if you need both

1. Test your platform

   b. **Spark_Pi.xml** to only test your Spark platform (compute PI)
   c. **Spark_Write_Read_HDFS.xml** to test both your HDFS and Spark platforms (write and read objects in Spark from/to the HDFS)

Swarm wkw params

Swarm

Description : Deployment of Docker Swarm. swarm_manager_port refers to the communication port of the swarm manager. consul_UI_port is the Consul web portal port. All started docker containers are prefixed with instance_name. network_name refers to the docker containers network name and subnet to the subnet mask.

Project name : Cloud Automation - Deployment

Bucket name : cloud-automation

Documentation : https://ow2-proactive.github.io/proactive-examples/DockerSwarm/resources/doc/V1/activeeon-deploy-swarm-hdfs-spark.pdf

consul_UI_port	8500	port of the consul web portal
swarm_manager_port	4000	port of the swarm manager process
instance_name	my-instance	base name of the docker containers (consul, swarm)
network_name	my-net	name of the docker network to connect
subnet	25.25.25.0/24	subnet of the docker network

HDFS wkw params

HDFS

Description : Deployment of HDFS.

Project name : Cloud Automation - Deployment

Bucket name : cloud-automation

Documentation : https://ow2-proactive.github.io/proactive-examples/DockerSwarm/resources/doc/V1/activeeon-deploy-swarm-hdfs-spark.pdf

HDFS_UI_port	6500	port of the HDFS web portal
datanode_starting_port	50010	Port range start of the datanodes
fs_name	25.25.25.2	name of the default file system
instance_name	my-instance	base name of the HDFS docker containers
network_name	my-net	name of the docker network to connect

Spark wkw params

Spark

Description : Deployment of Spark. spark_UI_port is the Spark web portal port. All started docker containers are prefixed with instance_name. network_name refers to the docker containers network name.

Project name : Cloud Automation - Deployment

Bucket name : cloud-automation

Documentation : https://ow2-proactive.github.io/proactive-examples/DockerSwarm/resources/doc/V1/activeeon-deploy-swarm-hdfs-spark.pdf

spark_UI_port	5000	port of the Spark web portal
instance_name	my-instance	base name of the HDFS docker containers
network_name	my-net	name of the docker network to connect

Spark_Pi wkw params

Spark_Pi

Description : A workflow to submit a Spark job from a docker container, to estimate Pi. This workflow requires a Spark platform.

Project name : Basic Big Data

Bucket name : big-data

Documentation : https://ow2-proactive.github.io/proactive-examples/DockerSwarm/resources/doc/V1/activeeon-deploy-swarm-hdfs-spark.pdf

slices	100
service_instance_id	1

- number of slices to cut the random number dataset into. Spark will run one task for each slice of the clusterport of the Spark web portal
- instance id of your cloud automation service

Spark_Write_Read_HDFS wkw params

Spark_Write_Read_HDFS

Description : A workflow to submit a Spark job from a docker container, to read/write files from/to HDFS. This workflow requires a Spark/HDFS platform.

Project name : Basic Big Data

Bucket name : big-data

Documentation : https://ow2-proactive.github.io/proactive-examples/DockerSwarm/resources/doc/V1/activeeon-deploy-swarm-hdfs-spark.pdf

service_instance_id	1	instance id of your cloud automation service
parquet_file_path	/user/hdfs/wiki/testwiki	parquet file path on the hdfs
csv_file_path	/user/hdfs/wiki/testwiki.csv	csv file path on the hdfs