Running Apache Spark in a Docker Container

Overview

The following instructions can be used to install Spark in a SLES 12 SP1/Docker instance running on IBM LinuxONE Community Cloud. One example provides an interactive Spark shell to store and read values from an array. A second example starts Spark master and worker nodes to estimate the value of Pi.

Instructions

1. Switch to root user for this exercise.
   > sudo -i

   ```
   linux1@dockerspark2:~> sudo -i
   dockerSpark2:~ #
   ```

2. Start the Docker service on your Linux system.
   # systemctl start docker

   ```
   dockerSpark2:~ # systemctl start docker
   dockerSpark2:~ #
   ```

3. Download the Spark Docker file from the repository.
   # docker pull brunswickheads/spark-1.5.2-s390x

   ```
   dockerSpark2:/# docker pull brunswickheads/spark-1.5.2-s390x
   Using default tag: latest
   latest: Pulling from brunswickheads/spark-1.5.2-s390x
   b5257eca2dd8: Pull complete
   78745e6c93f9: Pull complete
   b6a646b67400: Pull complete
   b09ba56b8f32: Pull complete
   ca77adadce6: Pull complete
   a8e5f3181b12: Pull complete
   Digest: sha256:c55f9185e571fbbccd6a2a2c126aed51700a441c67c8368da24b060612c0be7a
   Status: Downloaded newer image for brunswickheads/spark-1.5.2-s390x:latest
   ```

4. Launch a Spark Docker container and open the ports required for Spark. Start a bash shell within the container.
   # docker run -p 8080:8080 -p 8081:8081 -p 7077:7077 -p 6066:6066 -i -t brunswickheads/spark-1.5.2-s390x /bin/bash

   ```
   dockerSpark2:/# docker run -p 8080:8080 -p 8081:8081 -p 7077:7077 -p 6066:6066 -i -t brunswickheads/spark-1.5.2-s390x /bin/bash
   bash-4.2$
   ```

5. Exit the bash shell.
   # exit

© 2016 IBM Corporation
Running Apache Spark in a Docker Container

bash-4.2# exit
exit
dockerspark2:/ #

6. The docker container is stopped when you exit the bash shell.
   # docker ps

7. Switch to the Docker containers folder.
   # cd /data/docker/containers/

8. List the contents of the folder to view your Docker containers.
   # ls

9. Restart your Docker container using the first three characters of your Docker container directory.
   # docker start 422

10. Check the status of your running Docker container.
    # docker ps
    Note your Docker container id.

11. Attach to your running Docker container to restart the bash shell.
    a) # docker attach 422
    b) Hit ‘enter’ or ‘return’ again

12. Start an interactive Spark shell.
    # /opt/ibm/spark/bin/spark-shell
13. Test the Spark read-eval-print loop (REPL).
   a) Create a collection of 20 entries
   ```scala
   scala> val data = 1 to 20
   collection.immutable.Range.Inclusive = Range(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20) data: scala.Collection[Int] = Range(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20)
   ```

   b) Create a resilient distributed dataset (RDD)
   ```scala
   scala> val distdata = sc.parallelize(data)
   distdata: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[0] at parallelize at <console>:23
   ```

   c) Select and return the even values
   ```scala
   scala> distdata.filter(_ % 2 == 0).collect()
   res0: Array[Int] = Array(2, 4, 6, 8, 10, 12, 14, 16, 18, 20)
   ```

14. Exit from the Spark REPL.
   ```scala
   scala> :q
   ```
15. Switch to the Spark installation directory.
   # cd opt/ibm/spark/

16. Start the Spark Master server.
   # ./sbin/start-master.sh

17. Start a Spark Worker in your Docker container and allocate 2GB of memory. Use your Docker container id.
   # ./sbin/start-slave.sh -m 2G spark://<Docker container id>:7077

18. Using a browser, go to <your Linux ip address>:8080 to verify the Spark GUI is operational.

19. Run an example scala program that estimates the value of Pi. Use your Docker container id.
   # ./bin/spark-submit --class org.apache.spark.examples.SparkPi --master spark://<Docker container id>:6066 --deploy-mode cluster /opt/ibm/spark/lib/spark-examples-1.5.2-hadoop2.6.0.jar
20. Using a browser, go to <your Linux ip address>:8081 to see the output. Click on the stdout link for your DriverID.

21. Verify the result.

22. Exit the bash shell.
   # exit

23. The docker container is now stopped.
   # docker ps
Running Apache Spark in a Docker Container

References

Links
https://docs.docker.com/engine/userguide/usingdocker/
https://hub.docker.com/r/brunswickheads/spark-1.5.2-s390x/
https://spark.apache.org/docs/latest/quick-start.html