Integrating R environment with Spark cluster using docker

using

ANIL KUMAR JAIN

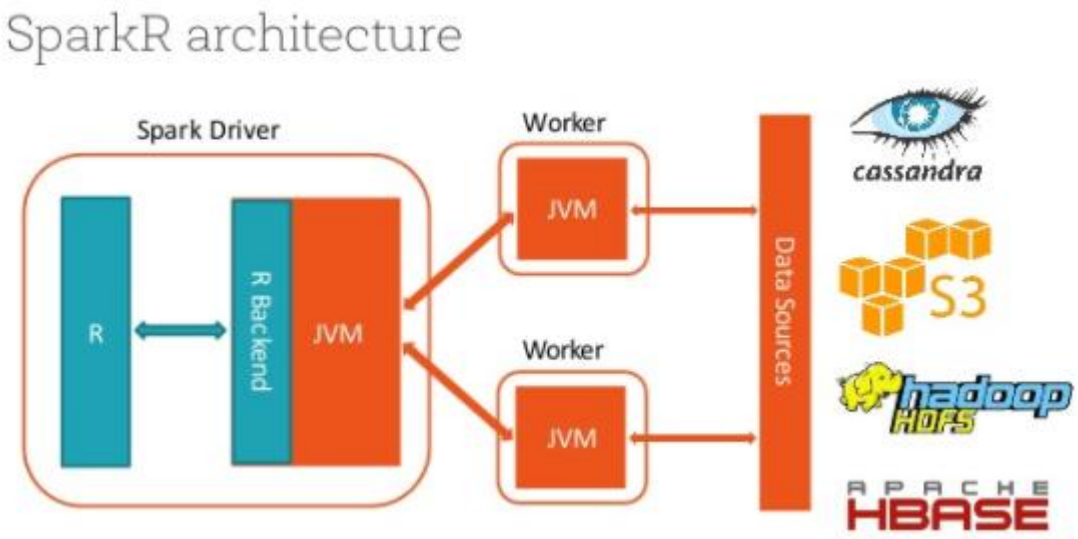
Topics

- Why Scalable machine learning environment with Spark cluster?
- Spark R architecture
- Installation Steps for Integrating R environment with Spark cluster
- A sample model execution in Spark R environment
- Summary

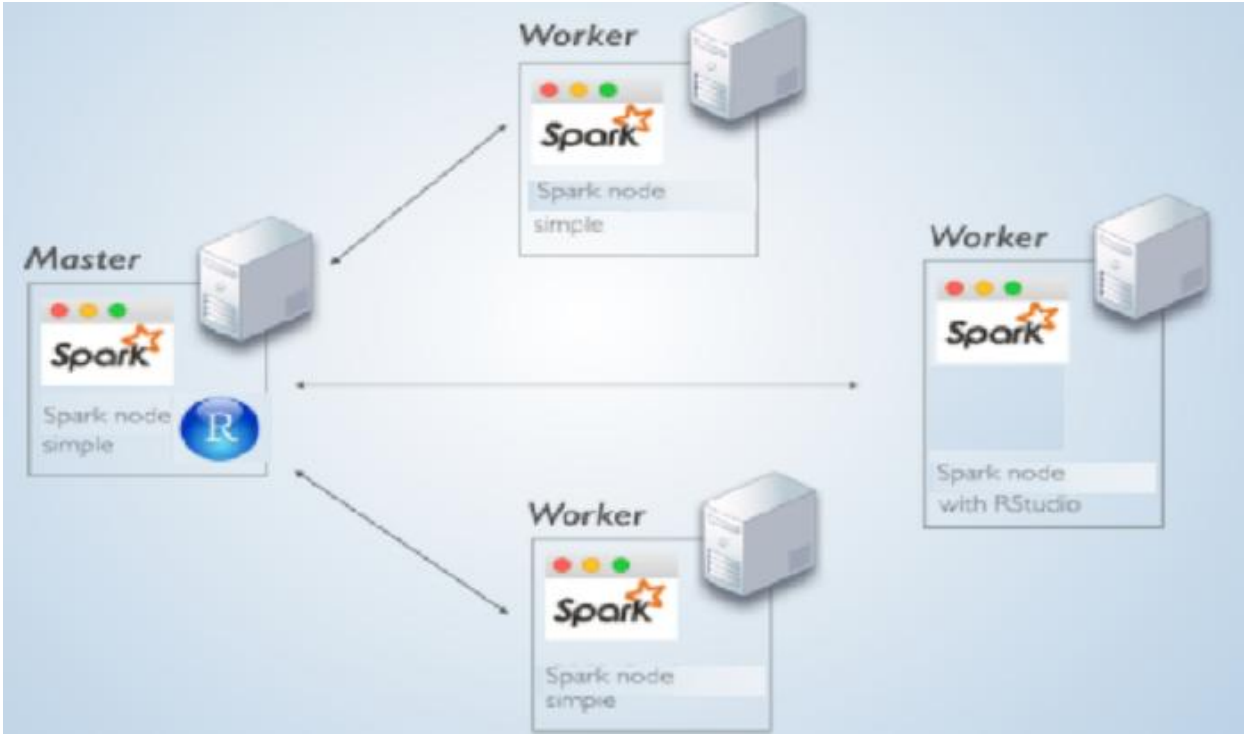
Need for Scalable machine learning environment

- R Studio/programming language is a machine learning tool for data scientist to develop, write and build machine learning projects. Alone R environment is not s ufficient to scale up for large data computation.
- Also in the production environment, data scientists have to run their model on th e actual data set which are there in the data lake or data warehouse. In absence of proper data pipeline infrastructure, it will make the make the life of data scienti st difficult to execute and deploy the model on the actual dataset.
- Here comes the need for scalable machine learning environemnt and there can not be better solution than integrating R environemnt with Spark cluster. It makes sure that data is seamlessly available to run the machine learning models and al gorithm on the actual data.

Spark R architecture



R Studio server setup with Spark cluster - diagram



Installation Steps for Integrating R environment with Spark cluster

```
## Step 1: Get the docker image for SparkR set up
docker pull angelsevillacamins/spark-rstudio-shiny
## Step 2: Define a network
docker network create spark_network
## Step3: Create data volume container with a folder to share among the nodes
docker create --net spark_network --name data-share
  --volume /home/rstudio/share angelsevillacamins/spark-rstudio-shiny
## Step 4: Deploy master node
docker run -d --net spark_network --name master  -p 8080:8080 -p 8787:8787 -p 80:3838  --volumes-from data-share    --restart=always ₩
  angelsevillacamins/spark-rstudio-shiny /usr/bin/supervisord --configuration=/opt/conf/master.conf
## Step 5: Changing permissions in the share folder of the data volume
docker exec master chmod a+w /home/rstudio/share
## Step 6: Deply worker01 node
docker run -d --net spark_network --name worker01 --volumes-from data-share  --restart=always
  angelsevillacamins/spark-rstudio-shiny /usr/bin/supervisord --configuration=/opt/conf/worker.conf
## Step 7:  Changing permissions in the share folder of the data volume
docker exec worker01 chmod a+w /home/rstudio/share
## Step 8: Deploy worker02 node
docker run -d --net spark_network --name worker02  --volumes-from data-share  --restart=always ₩
  angelsevillacamins/spark-rstudio-shiny /usr/bin/supervisord --configuration=/opt/conf/worker.conf
## Step 9: Changing permissions in the share folder of the data volume
docker exec worker02 chmod a+w /home/rstudio/share
```

If all goes well, Spark server and R studio server are available on web browser
Spark server -> http://your.ip.as.above:8080.
R Studio server -> , thus, http://your.ip.as.above:8787

Summary

- The need for scalable machine learning environemnt is must for data scientists for their machine learning algorithm to learn on large data set for better performance.
- There can not be better solution than integrating R environemnt with Spark cluster

links

- https://spark.rstudio.com/
- https://blog.rstudio.com/2016/09/27/sparklyr-r-interface-for-apache-spark/
- https://blog.rstudio.com/categories/packages
- https://spark.rstudio.com/examples-cdh.html
- https://acadgild.com/blog/how-to-integrate-r-with-spark/