Docker Spark Training

Kirk S. Kalvar

# Overview

Attended an evening Meetup introduction to Spark. The company that provide the presentation provided a Vagrant VM that no one could get to run. So I setup a Docker file to build the basic demo.

# Prerequisites

Installed DockerToolbox-1.9.1f for Windows

https://github.com/docker/toolbox/releases

Note: DockerToolbox-1.9.1f for Windows contains boot2docker.iso v1.9.1 but has a bug which will keep java from being installed. You need to download the v1.9.0.

boot2docker.iso v1.9.0

https://github.com/boot2docker/boot2docker/releases

# Download tetra-spark-training.tar.gz

https://drive.google.com/file/d/0B54qWs-0SiLNUXBCaENRX1JZUUk/view?usp=sharing

# Building Docker Container

cd DockerSparkTraining

cp <downloaded location>/tetra-spark-training.tar.gz .

docker build -t tetra-spark-training .

# Running the Docker Container

docker run -it tetra-spark-training bash

# Running the Tetra Spark Training Demo "inside the container"

cd /tetra-spark-training

./spark.sh

# Testing the Sparking Training Demo

Once spark starts, cut and paste the following into the spark shell:

import org.apache.spark.SparkContext.\_

import org.apache.spark.rdd.\_

import com.uebercomputing.mailparser.enronfiles.AvroMessageProcessor

import com.uebercomputing.mailrecord.\_

import com.uebercomputing.mailrecord.Implicits.mailRecordToMailRecordOps

val args = Array("--avroMailInput", "../../data/filemail.avro", "--hadoopConfPath", "hadoop-local.xml")

val config = CommandLineOptionsParser.getConfigOpt(args).get

val recordsRdd = MailRecordAnalytic.getMailRecordsRdd(sc, config)

recordsRdd.count # should output the number "Long = 40419"