RHadoop InstallationGuide

Ha Nguyen

June 25, 2015

Goals

Install RHadoop system for testing R capability to manage and analyze data in Hadoop cluster

Components

- Operating System: Ubuntu Server 14.04 LTS(HVM)
- Apache Hadoop 2.7.0 Single Node Cluster
- R & Rstudio Server
- RHadoop packages:

Installation Steps:

1. Install Ubuntu Server on Amazon EC2

- Ubuntu Server 14.04 LTS(HVM) t2.micro (Variable ECUs, 1 vCPUs, 2.5 GHz, Intel Xeon Family, 1 GiB memory, EBS only) *free tier eligible*
- http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-launch-instance_linux.html

2. Setup Ubuntu system

- Connnect to Ubuntu system with ssh using private keypair
- Increase swapfile to overcome low Ram capacity: https://github.com/wcyuan/data-science-courses

3. Set up Apache Hadoop 2.7.0 Single Node Cluster

- https://rstudio-pubs-static.s3.amazonaws.com/78508_abe89197267240dfb6f4facb361a20ed.html
- Make sure nodes be able to connect to each other with ssh using keypair

3. Set up R:

- Update R repository in sources.list file to obtain latest R packages http://cran.r-project.org/bin/linux/ubuntu/

4. Set up Rstudio Server

- http://www.rstudio.com/products/rstudio/download-server/

5. Install RHadoop packages

- Install pre-required R packages:

- Rcpp", "RJSONIO", "bitops", "digest", "functional", "stringr", "plyr", "reshape2", "dplyr", "R.methodsS3", "caTools", "Hmisc", "rjson", "memoise", "data.table", "rJava"
- Set up environment variables:
- Sys.setenv("HADOOP_CMD"="/usr/local/hadoop/bin/hadoop")
- Sys.setenv("LD_LIBRARY_PATH"="/usr/local/hadoop/lib/native/")
- Download RHadoop packages: https://github.com/RevolutionAnalytics/RHadoop/wiki
- Install Rhadoop packages:

```
install.packages("<path>/rhdfs*.tar.gz", repos=NULL, type="source")
install.packages("<path>/rmr2*.tar.gz", repos=NULL, type="source")
install.packages("<path>plyrmr*.tar.gz", repos=NULL, type="source")
```

Testing

1. Test Hadoop MapReduce job with example

- Calculating *pi*

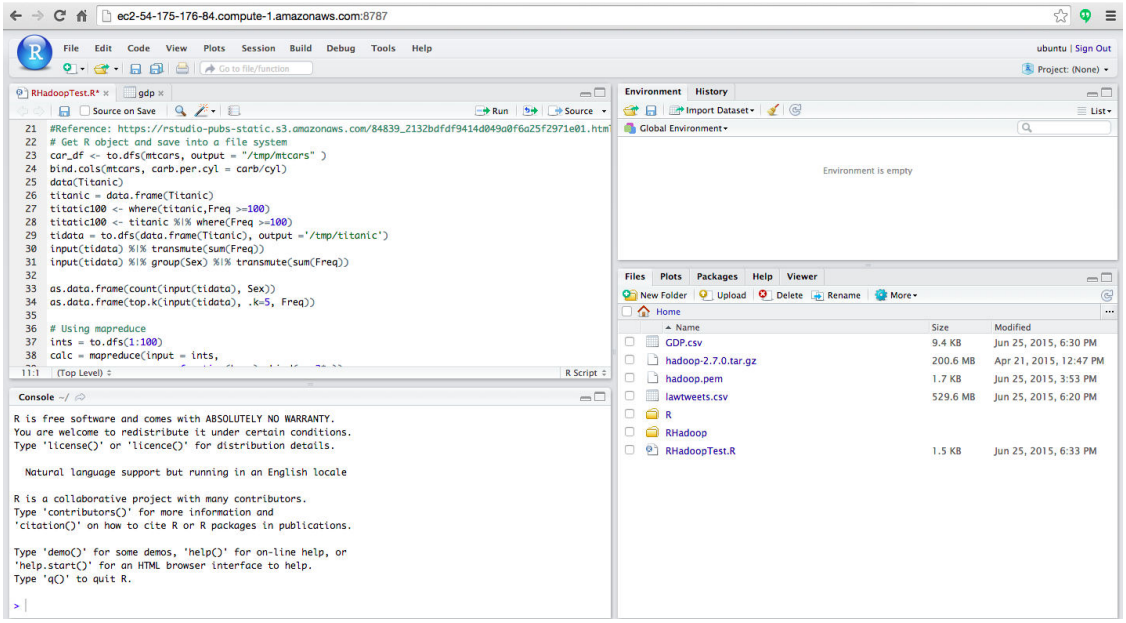
$ cd $HADOOP_COMMON_HOME

$ bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.0.jar pi 10 100

```
Number of Maps  = 10
Samples per Map = 100
Wrote input for Map #0
Wrote input for Map #1
Wrote input for Map #2
Wrote input for Map #3
Wrote input for Map #4
Wrote input for Map #5
Wrote input for Map #6
Wrote input for Map #7
Wrote input for Map #8
Wrote input for Map #9
Starting Job
15/06/26 13:26:10 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
15/06/26 13:26:10 INFO input.FileInputFormat: Total input paths to process : 10
15/06/26 13:26:10 INFO mapreduce.JobSubmitter: number of splits:10
15/06/26 13:26:10 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1435325139338_0001
15/06/26 13:26:11 INFO impl.YarnClientImpl: Submitted application application_1435325139338_0001
15/06/26 13:26:11 INFO mapreduce.Job: The url to track the job: http://ec2-54-175-176-84.compute-1.ama
zonaws.com:8088/proxy/application_1435325139338_0001/
15/06/26 13:26:11 INFO mapreduce.Job: Running job: job_1435325139338_0001
15/06/26 13:26:21 INFO mapreduce.Job: Job job_1435325139338_0001 running in uber mode : false
15/06/26 13:26:21 INFO mapreduce.Job:  map 0% reduce 0%
15/06/26 13:27:14 INFO mapreduce.Job:  map 20% reduce 0%
15/06/26 13:27:16 INFO mapreduce.Job:  map 60% reduce 0%
15/06/26 13:27:53 INFO mapreduce.Job:  map 100% reduce 20%
15/06/26 13:27:55 INFO mapreduce.Job:  map 100% reduce 100%
15/06/26 13:27:56 INFO mapreduce.Job: Job job_1435325139338_0001 completed successfully
```

2. Test R and Rstudio Server

- Using webbrowser to connect to Rstudio Server: http://ec2-54-175-176-84.compute-1.amazonaws.com:8787/
- Username: ubuntu | Password: ubuntu

- 

3. Demo with RHadoop packages

4. Fixing Error when using Hadoop

- Can not connect to Server: http://77-thoughts.com/hadoop-info-ipc-client-retrying-connect-to-server-localhost127-0-0-19000/