

INFSCI 2750: Cloud Computing

Mini Project 2

README

1. Setting up Spark on YARN in remote Ubuntu server

Following picture is the screenshot of working system

The screenshot shows a terminal window on the left and a 'New Remote Connection' dialog on the right.

Terminal Window Content:

```
Welcome to Ubuntu 16.04.1 LTS (GNU/Linux 4.4.0-62-generic x86_64)

* Documentation: https://help.ubuntu.com
* Management: https://landscape.canonical.com
* Support: https://ubuntu.com/advantage

Get cloud support with Ubuntu Advantage Cloud Guest:
http://www.ubuntu.com/business/services/cloud

28 packages can be updated.
0 updates are security updates.

*** System restart required ***
Last login: Thu Mar 16 22:42:31 2017 from 104.236.234.130
[root@master:~# jps
10418 SecondaryNameNode
11362 Master
12121 Jps
10203 NameNode
10572 ResourceManager
[root@master:~# ssh slave
Welcome to Ubuntu 16.04.1 LTS (GNU/Linux 4.4.0-62-generic x86_64)

* Documentation: https://help.ubuntu.com
* Management: https://landscape.canonical.com
* Support: https://ubuntu.com/advantage

Get cloud support with Ubuntu Advantage Cloud Guest:
http://www.ubuntu.com/business/services/cloud

35 packages can be updated.
7 updates are security updates.

*** System restart required ***
Last login: Thu Mar 16 22:47:43 2017 from 104.236.235.155
[root@slave:~# jps
29441 DataNode
30593 Jps
30053 Worker
29575 NodeManager
root@slave:~# ]
```

New Remote Connection Dialog:

Service	Server
Secure Shell (ssh)	▶ 104.236.234.130
Secure File Transfer (sftp)	▶ 104.236.235.155
File Transfer (ftp)	▶ --- Discovered Servers ---
Remote Login (telnet)	▶ Judith's MacBook Pro
	▶ Slayers MacBook
	▶ tangsong的MacBook Pro
	▶ UPTV-2

Buttons at the bottom: + - (left), + - (right), User: root, SSH (Automatic), ssh root@104.236.235.155, Connect.

in Master server, there is Master.

In Slave server, there is a Worker

2. Develop spark program

(1) Using cached RDD operation to analyze `user_artists.dat` in standalone Spark
place the data file in the local working directory for example like `/opt/spark/spark-2.0.2-bin-hadoop2.7`

#start the spark by spark-shell

```
#input the data file in the spark RDD by sc
val data=sc.textFile("user_artists.dat").cache()
data.count
it should show 92834
```

```
#map the result
val mapresult=data.map(line=>line.split("\s+"))
mapresult.collect
```

```
#map the key-value pair
val mapkey=mapresult.map(arr => (arr(1),arr(2).toInt)).cache()
macky.collect
```

```
#reduce the result by key
val reduceresult=mapkey.reduceByKey(new org.apache.spark.HashPartitioner(2), (x,y)
=> x+y)
reduceresult.collect
#or
val reduceresult=mapkey.reduceByKey(_+_)
```

#sort the value descending

```
val sortresult=reduceresult.sortBy(_._2, false)
val final=sortresult.take(10)
```

the top 10 artists ID are 289, 72, 89, 292, 498, 67, 288, 701, 227, 300

following picture show the final result of top 10 artists and their listening counts.

```
scala> sortresult.take(10)
res11: Array[(String, Int)] = Array((289,2393140), (72,1301308), (89,1291387), (292,1058405), (498,963449), (67,921198), (288,905423), (701,688529), (227,662116), (300,532545))
```

```
scala> drwx|
```

(2) how to execute the .scala program.(sbt package tool)

- a. make sure that spark and scala are both implemented successfully in root user
- b. download sbt(simple buildtool) which is used for package the scala file and produce the .jar file. Modify the ~/.bashrc by adding path variable and install the sbt. Checking the whether sbt is working, by sbt sbt-version.
- c. Create a scala file directory, for example “mkdir –p ./spark/src/main/scala” and create program file under the “scala” file. The file structure should look like this

```
./src  
./src/main  
./src/main/scala  
./src/main/scala/SimpleApp.scala  
./src/main/scala/SimpleApp.scala~  
root@song-VirtualBox:~/spark#
```

and then use “sbt package” command to package this scala program file into jar file, which is very time consuming , it will take approximate 200 second to finish the package.

And then in the spark file directory, enter “bin” file and command “spark-submit”

Enter the flowing command in the terminal

“spark-submit --class "SimpleApp" ~/spark/target/scala-2.11/simple-project_2.11-1.0.jar” and it will show the result or all the result will be saved to the new specified output file.

Following picture is the running result for the question 3(1) and (2)

```
17/03/22 18:41:44 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndPoint: OutputCommitCoordinator stopped!  
17/03/22 18:41:44 INFO SparkContext: Successfully stopped SparkContext  
17/03/22 18:41:44 INFO ShutdownHookManager: Shutdown hook called  
17/03/22 18:41:44 INFO ShutdownHookManager: Deleting directory /tmp/spark-05801d12-d9da-48c0-a603-8db823332a0f  
root@song-VirtualBox:/opt/spark/spark-2.0.2-bin-hadoop2.7/bin# spark-submit --class "SimpleApp" ~/spark/target/scala-2.11/simple-project_2.11-8.0.jar 2>%1 | grep "there are"  
there are 293 hits in /assests/js/lorpro.js and there are 66831 in /favicon.ico
```

3. Develop Spark Program

Cached RDD method:

(1) there are 293 hit for the /assets/js/lowpro.js

```
val logfile=sc.textFile("access_log.dat").cache()  
val mapkey=logfile.map(word=>word.split("\\s+"))  
val mapfilter=mapkey.filter(line=>line.contains("/assets/js/lowpro.js")).cache()  
val result=mapfilter.count()
```

```
scala> logfile.count  
res30: Long = 4477843  
  
scala> val result=logfile.map(word=>word.split("\\s+")).filter(line=>line.contains("/as-  
sets/js/lowpro.js")).count()  
result: Long = 293  
  
scala>
```

(2) there are 66831 record for "/favicon.ico"

```
val logfile=sc.textFile("access_log.dat").cache()  
val mapkey=logfile.map(word=>word.split("\\s+"))  
val mapfilter=mapkey.filter(line=>line.contains("/favicon.ico"))  
val result=mapfilter.count()
```

```
scala> val result=logfile.map(word=>word.split("\\s+")).filter(line=>line.  
contains("favicon.ico")).count()  
result: Long = 66831  
  
scala>
```

(3) the website path have the most hits number is “/assets/css/combined.css” there are 117348 hits for this site path.

#RDD operation and transition

```
val logfile=sc.textFile("access_log.dat").cache()
reduceresult=logfile.map(line=>line("\s+")).cache().map(arr=>(arr(6),1)).reduceByKey(_+_).sortBy(_._2, false).take(1)
```

```
scala> sortresult.collect
res20: Array[(String, Int)] = Array((/assets/css/combined.css,117348), (/assets/js/java
script_combined.js,106818), (/99303), (/assets/img/home-logo.png,98744), (/assets/css/
printstyles.css,93158), (/images/filmpics/0000/3695/Pelican_Blood_2D_Pack.jpg,91933), (
/favicon.ico,66831), (/robots.txt,51975), (/images/filmpics/0000/3139/SBX476_Vanquisher_
2d.jpg,39591), (/assets/img/search-button.gif,38990), (/assets/img/play_icon.png,34151),
(/images/filmmediablock/290/Harpoon_2d.JPG,32533), (/assets/img/x.gif,29377), (/imag
es/filmpics/0000/1421/RagingPhoenix_2DSleeve.jpeg,29243), (/release-schedule/,25937), (
/assets/img/release-schedule-logo.png,24292), (/search/,23055), (/assets/img/banner/ten-
years-banner-grey.jpg,22129), (/assets/img/banner/ten-years-banner-white.jpg,22121), (
/assets/img/ba...
scala> sortresult.take(1)
res21: Array[(String, Int)] = Array((/assets/css/combined.css,117348))
```

(4) the IP address have the most access website is “10.216.113.172”. there are 158614 access records for this IP address

#RDD operation and transition

```
val logfile=sc.textFile("access_log.dat").cache()
val reduceresult=logfile.map(line=>line("\s+")).map(arr=>(arr(0),1)).reduceByKey(_+_).
sortBy(_._2, false).take(1)
```

```
scala> reduceresult.collect
res23: Array[(String, Int)] = Array((10.216.113.172,158614),
173.141.213.47503), (10.240.144.183,43592), (10.41.69.177,375),
(10.211.47.159,20866), (10.96.173.111,19667), (10.203.77.1
721), (10.118.250.30,18282), (10.56.48.40,17850), (10.194.74.
,16709), (10.50.199.54,15717), (10.247.111.104,15583), (10.10
4.212,14420), (10.152.195.138,13775), (10.238.101.239,12450),
10.25.20.51,12074), (10.216.227.195,11975), (10.118.19.97,115),
(10.1.181.142,10686), (10.39.94.109,10681), (10.54.143.175
23), (10.38.159.247,9773), (10.116.44.213,9180), (10.179.244.
8722), (10.143...
scala> reduceresult.take(1)
<console>:24: error: not found: value reduceresult
      reduceresult.take(1)
      ^

scala> reduceresult.take(1)
res25: Array[(String, Int)] = Array((10.216.113.172,158614))
```

4. Time Performance, running with cache RDD

```
Val logfile=sc.textFile("access_log.dat").cache()
```

Completed Stages: 3

Completed Stages (3)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read
2	take at <console>:34 +details	2017/03/23 00:38:05	0.7 s	1/1			274.7 KB
1	sortBy at <console>:32 +details	2017/03/23 00:38:04	0.9 s	1/1			114.8 KB
0	map at <console>:28 +details	2017/03/23 00:38:01	2 s	1/1	1266.0 KB		

Time Performance(running without the cache RDD) spend more time.

Completed Stages: 33

Completed Stages (33)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
45	count at <console>:29 +details	2017/03/22 12:45:55	12 s	15/15	482.4 MB			
44	count at <console>:29 +details	2017/03/22 12:43:58	12 s	15/15	482.4 MB			
43	count at <console>:29 +details	2017/03/22 12:43:27	12 s	15/15	482.4 MB			
42	collect at <console>:29 +details	2017/03/22 12:42:46	12 s	15/15	482.4 MB			
41	take at <console>:29 +details	2017/03/22 12:35:44	67 ms	1/1			329.3 KB	
38	collect at <console>:29 +details	2017/03/22 12:34:51	2 s	15/15			4.4 MB	
37	sortBy at <console>:26 +details	2017/03/22 12:34:50	1 s	15/15			3.3 MB	4.4 MB
35	sortBy at <console>:26 +details	2017/03/22 12:34:39	0.6 s	15/15			3.3 MB	
34	count at <console>:29 +details	2017/03/22 12:34:38	1 s	15/15	482.4 MB			