Exerc?cios RDD com Parti??es

April 11, 2022

1 Exercícios RDD com Partições

1.1 1. Ler com RDD os arquivos localmente do diretório "/opt/spark/logs/" ("file:///opt/spark/logs/") com 10 partições

In [1]:
```
!ls /opt/spark/logs
```

```
spark--org.apache.spark.deploy.master.Master-1-jupyter-notebook.out
```

In [2]:
```
!cat //opt/spark/logs/spark--org.apache.spark.deploy.master.Master-1-jupyter-notebook.
```

```
Spark Command: /opt/java/bin/java -cp /etc/spark/:/opt/spark/jars/*:/etc/hadoop/:/etc/hadoop/*
========================================
20/03/18 20:31:09 INFO master.Master: Started daemon with process name: 1087@jupyter-notebook
20/03/18 20:31:09 INFO util.SignalUtils: Registered signal handler for TERM
20/03/18 20:31:09 INFO util.SignalUtils: Registered signal handler for HUP
20/03/18 20:31:09 INFO util.SignalUtils: Registered signal handler for INT
20/03/18 20:31:10 INFO spark.SecurityManager: Changing view acls to: root
20/03/18 20:31:10 INFO spark.SecurityManager: Changing modify acls to: root
20/03/18 20:31:10 INFO spark.SecurityManager: Changing view acls groups to:
20/03/18 20:31:10 INFO spark.SecurityManager: Changing modify acls groups to:
20/03/18 20:31:10 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls
20/03/18 20:31:10 INFO util.Utils: Successfully started service 'sparkMaster' on port 7077.
20/03/18 20:31:10 INFO master.Master: Starting Spark master at spark://localhost:7077
20/03/18 20:31:10 INFO master.Master: Running Spark version 2.4.1
20/03/18 20:31:10 INFO util.log: Logging initialized @1454ms
20/03/18 20:31:10 INFO server.Server: jetty-9.3.z-SNAPSHOT, build timestamp: unknown, git hash
20/03/18 20:31:10 INFO server.Server: Started @1504ms
20/03/18 20:31:10 INFO server.AbstractConnector: Started ServerConnector@5526ec85{HTTP/1.1,[ht
20/03/18 20:31:10 INFO util.Utils: Successfully started service 'MasterUI' on port 8080.
20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@78c62b96{
20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@1330e23c{
20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@3f7fb168{
20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@50e5c33c{
20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@44d0ea1e{
20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@2a362185{
20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@222db61{/
```

```
20/03/18 20:31:10 INFO ui.MasterWebUI: Bound MasterWebUI to 0.0.0.0, and started at http://jupy
20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@58cb287e{/
20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@3d2d7a89{/
20/03/18 20:31:10 INFO master.Master: I have been elected leader! New state: ALIVE
```

1.2 2. Com uso de RDD faça as seguintes operações

1.2.1 a) Contar a quantidade de linhas

```
In [3]: log = sc.textFile("file:///opt/spark/logs")
```

```
In [4]: log.count()
```

```
Out[4]: 30
```

1.2.2 b) Visualizar a primeira linha

```
In [5]: log.first()
```

```
Out[5]: 'Spark Command: /opt/java/bin/java -cp /etc/spark/:/opt/spark/jars/*:/etc/hadoop/:/etc/
```

1.2.3 c) Visualizar todas as linhas

```
In [6]: log.collect()
```

```
Out[6]: ['Spark Command: /opt/java/bin/java -cp /etc/spark/:/opt/spark/jars/*:/etc/hadoop/:/etc
         '========================================',
         '20/03/18 20:31:09 INFO master.Master: Started daemon with process name: 1087@jupyter-
         '20/03/18 20:31:09 INFO util.SignalUtils: Registered signal handler for TERM',
         '20/03/18 20:31:09 INFO util.SignalUtils: Registered signal handler for HUP',
         '20/03/18 20:31:09 INFO util.SignalUtils: Registered signal handler for INT',
         '20/03/18 20:31:10 INFO spark.SecurityManager: Changing view acls to: root',
         '20/03/18 20:31:10 INFO spark.SecurityManager: Changing modify acls to: root',
         '20/03/18 20:31:10 INFO spark.SecurityManager: Changing view acls groups to: ',
         '20/03/18 20:31:10 INFO spark.SecurityManager: Changing modify acls groups to: ',
         '20/03/18 20:31:10 INFO spark.SecurityManager: SecurityManager: authentication disable
         "20/03/18 20:31:10 INFO util.Utils: Successfully started service 'sparkMaster' on port
         '20/03/18 20:31:10 INFO master.Master: Starting Spark master at spark://localhost:7077
         '20/03/18 20:31:10 INFO master.Master: Running Spark version 2.4.1',
         '20/03/18 20:31:10 INFO util.log: Logging initialized @1454ms',
         '20/03/18 20:31:10 INFO server.Server: jetty-9.3.z-SNAPSHOT, build timestamp: unknown
         '20/03/18 20:31:10 INFO server.Server: Started @1504ms',
         '20/03/18 20:31:10 INFO server.AbstractConnector: Started ServerConnector@5526ec85{HTT
         "20/03/18 20:31:10 INFO util.Utils: Successfully started service 'MasterUI' on port 80
         '20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@
         '20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@
         '20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@
         '20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@
```

```
'20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler
'20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler
'20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler
'20/03/18 20:31:10 INFO ui.MasterWebUI: Bound MasterWebUI to 0.0.0.0, and started at
'20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler
'20/03/18 20:31:10 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler
'20/03/18 20:31:10 INFO master.Master: I have been elected leader! New state: ALIVE']
```

1.2.4 d) Contar a quantidade de palavras

```
In [9]: #Separar as palavras da linhas
        palavras = log.flatMap(lambda linha: linha.split(" "))
        palavras.count()

Out[9]: 268

In [10]: palavras.first()

Out[10]: 'Spark'
```

1.2.5 e) Converter todas as palavras em minúsculas

```
In [11]: minuscula = palavras.map(lambda palavra: palavra.lower())

In [12]: minuscula.first()

Out[12]: 'spark'
```

1.2.6 f) Remover as palavras de tamanho menor que 2

```
In [14]: minuscula_maior2 = minuscula.filter(lambda palavra: len(palavra) > 2)
         minuscula_maior2.count()

Out[14]: 257
```

1.2.7 g) Atribuir o valor de 1 para cada palavra

```
In [16]: palavra_1 = minuscula_maior2.map(lambda palavra: (palavra, 1))
         palavra_1.first()

Out[16]: ('spark', 1)
```

1.2.8 h) Contar as palavras com o mesmo nome

```
In [17]: palavra_reduce = palavra_1.reduceByKey(lambda chave1, chave2: chave1 + chave2)
         palavra_reduce.count()

Out[17]: 101
```

1.2.9 i) Visualizar em ordem alfabética

```
In [18]: palavra_ordem = palavra_reduce.sortBy(lambda palavra: palavra[0])
         palavra_ordem.collect()
```

```
Out[18]: [("'masterui'", 1),
          ("'sparkmaster'", 1),
          ('--host', 1),
          ('--port', 1),
          ('--webui-port', 1),
          ('-cp', 1),
          ('-xmx1g', 1),
          ('/etc/spark/:/opt/spark/jars/*:/etc/hadoop/:/etc/hadoop/*:/opt/hadoop/share/hadoop/c
           1),
          ('/opt/java/bin/java', 1),
          ('0.0.0.0,', 1),
          ('1087@jupyter-notebook', 1),
          ('2.4.1', 1),
          ('20/03/18', 28),
          ('20:31:09', 4),
          ('20:31:10', 24),
          ('7077', 1),
          ('7077.', 1),
          ('8080', 1),
          ('8080.', 1),
          ('=======================================', 1),
          ('@1454ms', 1),
          ('@1504ms', 1),
          ('acls', 5),
          ('alive', 1),
          ('and', 1),
          ('authentication', 1),
          ('been', 1),
          ('bound', 1),
          ('build', 1),
          ('changing', 4),
          ('command:', 1),
          ('daemon', 1),
          ('disabled;', 2),
          ('elected', 1),
          ('for', 3),
          ('git', 1),
          ('groups', 4),
          ('handler', 3),
          ('handler.contexthandler:', 9),
          ('hash:', 1),
          ('have', 1),
          ('http://jupyter-notebook:8080', 1),
          ('hup', 1),
```

```
('info', 28),
('initialized', 1),
('int', 1),
('jetty-9.3.z-snapshot,', 1),
('leader!', 1),
('localhost', 1),
('logging', 1),
('master', 1),
('master.master:', 4),
('masterwebui', 1),
('modify', 4),
('name:', 1),
('new', 1),
('o.s.j.s.servletcontexthandler@1330e23c{/app/json,null,available,@spark}',
 1),
('o.s.j.s.servletcontexthandler@222db61{/driver/kill,null,available,@spark}',
 1),
('o.s.j.s.servletcontexthandler@2a362185{/app/kill,null,available,@spark}',
 1),
('o.s.j.s.servletcontexthandler@3d2d7a89{/metrics/applications/json,null,available,@s
 1),
('o.s.j.s.servletcontexthandler@3f7fb168{/,null,available,@spark}', 1),
('o.s.j.s.servletcontexthandler@44d0ea1e{/static,null,available,@spark}', 1),
('o.s.j.s.servletcontexthandler@50e5c33c{/json,null,available,@spark}', 1),
('o.s.j.s.servletcontexthandler@58cb287e{/metrics/master/json,null,available,@spark}
 1),
('o.s.j.s.servletcontexthandler@78c62b96{/app,null,available,@spark}', 1),
('org.apache.spark.deploy.master.master', 1),
('permissions:', 4),
('port', 2),
('process', 1),
('registered', 3),
('root', 2),
('running', 1),
('securitymanager:', 1),
('server.abstractconnector:', 1),
('server.server:', 2),
('serverconnector@5526ec85{http/1.1,[http/1.1]}{0.0.0.0:8080}', 1),
('service', 2),
('set()', 1),
('set();', 1),
('set(root);', 2),
('signal', 3),
('spark', 3),
('spark.securitymanager:', 5),
('spark://localhost:7077', 1),
('started', 15),
('starting', 1),
```

```
 ('state:', 1),
 ('successfully', 2),
 ('term', 1),
 ('timestamp:', 1),
 ('to:', 4),
 ('ui.masterwebui:', 1),
 ('unknown', 1),
 ('unknown,', 1),
 ('users', 2),
 ('util.log:', 1),
 ('util.signalutils:', 3),
 ('util.utils:', 2),
 ('version', 1),
 ('view', 4),
 ('with', 5)]
```

1.2.10 j) Visualizar em ordem decrescente a quantidade de palavras

```
In [22]: palavra_ordem_qtd = palavra_reduce.sortBy(lambda palavra: -palavra[1])
         palavra_ordem_qtd.collect()
```

```
Out[22]: [('20/03/18', 28),
          ('info', 28),
          ('20:31:10', 24),
          ('started', 15),
          ('handler.contexthandler:', 9),
          ('with', 5),
          ('spark.securitymanager:', 5),
          ('acls', 5),
          ('20:31:09', 4),
          ('master.master:', 4),
          ('groups', 4),
          ('permissions:', 4),
          ('changing', 4),
          ('view', 4),
          ('to:', 4),
          ('modify', 4),
          ('registered', 3),
          ('spark', 3),
          ('util.signalutils:', 3),
          ('signal', 3),
          ('handler', 3),
          ('for', 3),
          ('root', 2),
          ('disabled;', 2),
          ('set(root);', 2),
          ('successfully', 2),
          ('service', 2),
```

```
('server.server:', 2),
('users', 2),
('util.utils:', 2),
('port', 2),
('/opt/java/bin/java', 1),
('7077', 1),
('--webui-port', 1),
('8080', 1),
('daemon', 1),
('process', 1),
('term', 1),
('int', 1),
('securitymanager:', 1),
('set();', 1),
('set()', 1),
("'sparkmaster'", 1),
('starting', 1),
('master', 1),
('version', 1),
('2.4.1', 1),
('util.log:', 1),
('logging', 1),
('@1454ms', 1),
('jetty-9.3.z-snapshot,', 1),
('timestamp:', 1),
('git', 1),
('hash:', 1),
('unknown', 1),
('o.s.j.s.servletcontexthandler@1330e23c{/app/json,null,available,@spark}',
 1),
('o.s.j.s.servletcontexthandler@44d0ea1e{/static,null,available,@spark}', 1),
('masterwebui', 1),
('o.s.j.s.servletcontexthandler@3d2d7a89{/metrics/applications/json,null,available,@s
 1),
('have', 1),
('leader!', 1),
('new', 1),
('state:', 1),
('command:', 1),
('-cp', 1),
('/etc/spark/:/opt/spark/jars/*:/etc/hadoop/:/etc/hadoop/*:/opt/hadoop/share/hadoop/c
 1),
('-xmx1g', 1),
('org.apache.spark.deploy.master.master', 1),
('--host', 1),
('localhost', 1),
('--port', 1),
('========================================', 1),
```

```
 ('name:', 1),
 ('1087@jupyter-notebook', 1),
 ('hup', 1),
 ('authentication', 1),
 ('7077.', 1),
 ('spark://localhost:7077', 1),
 ('running', 1),
 ('initialized', 1),
 ('build', 1),
 ('unknown,', 1),
 ('@1504ms', 1),
 ('server.abstractconnector:', 1),
 ('serverconnector@5526ec85{http/1.1,[http/1.1]}{0.0.0.0:8080}', 1),
 ("'masterui'", 1),
 ('8080.', 1),
 ('o.s.j.s.servletcontexthandler@78c62b96{/app,null,available,@spark}', 1),
 ('o.s.j.s.servletcontexthandler@3f7fb168{/,null,available,@spark}', 1),
 ('o.s.j.s.servletcontexthandler@50e5c33c{/json,null,available,@spark}', 1),
 ('o.s.j.s.servletcontexthandler@2a362185{/app/kill,null,available,@spark}',
  1),
 ('o.s.j.s.servletcontexthandler@222db61{/driver/kill,null,available,@spark}',
  1),
 ('ui.masterwebui:', 1),
 ('bound', 1),
 ('0.0.0.0,', 1),
 ('and', 1),
 ('http://jupyter-notebook:8080', 1),
 ('o.s.j.s.servletcontexthandler@58cb287e{/metrics/master/json,null,available,@spark}
  1),
 ('been', 1),
 ('elected', 1),
 ('alive', 1)]
```

1.2.11 k) Remover as palavras, com a quantidade de palavras > 1

```
In [26]: palavra_ordem_filtro = palavra_ordem_qtd.filter(lambda palavra: palavra[1] > 1)
         palavra_ordem_filtro.collect()
```

```
Out[26]: [('20/03/18', 28),
          ('info', 28),
          ('20:31:10', 24),
          ('started', 15),
          ('handler.contexthandler:', 9),
          ('with', 5),
          ('spark.securitymanager:', 5),
          ('acls', 5),
          ('20:31:09', 4),
          ('master.master:', 4),
```

```
('groups', 4),
('permissions:', 4),
('changing', 4),
('view', 4),
('to:', 4),
('modify', 4),
('registered', 3),
('spark', 3),
('util.signalutils:', 3),
('signal', 3),
('handler', 3),
('for', 3),
('root', 2),
('disabled;', 2),
('set(root);', 2),
('successfully', 2),
('service', 2),
('server.server:', 2),
('users', 2),
('util.utils:', 2),
('port', 2)]
```

1.2.12 l) Salvar o RDD no diretorio do HDFS /user//logs_count_word

```
In [27]: palavra_ordem_filtro.saveAsTextFile("/user/jherfson/logs_count_word")
```

```
In [28]: !hdfs dfs -ls /user/jherfson/logs_count_word
```

```
Found 3 items
-rw-r--r--   2 root supergroup          0 2022-04-11 15:10 /user/jherfson/logs_count_word/_SUC
-rw-r--r--   2 root supergroup        517 2022-04-11 15:10 /user/jherfson/logs_count_word/part
-rw-r--r--   2 root supergroup          0 2022-04-11 15:10 /user/jherfson/logs_count_word/part
```