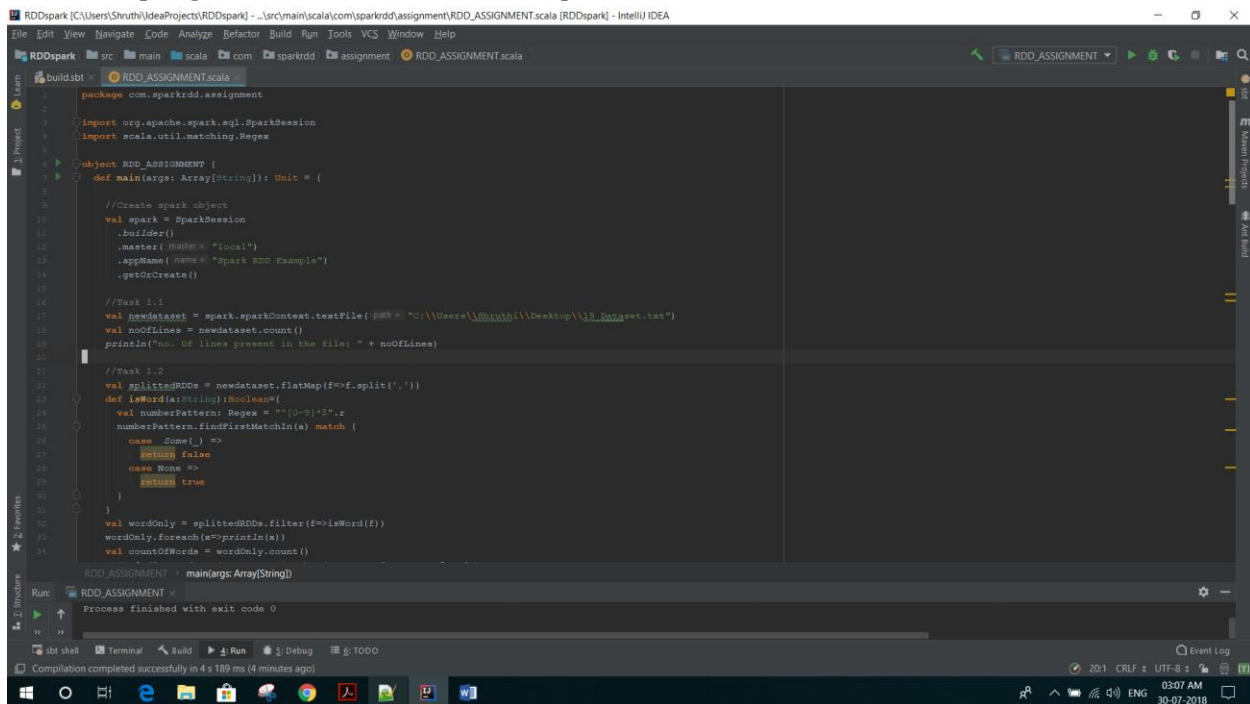


# RDD DEEP DIVE

Please find the code executed & results in screenshots, Thank you..!

## Task 1

1. Write a program to read a text file and print the number of rows of data in the document.



```
package com.sparkrdd.assignment

import org.apache.spark.sql.SparkSession
import scala.util.matching.Regex

object RDD_ASSIGNMENT {
  def main(args: Array[String]): Unit = {
    // Create spark object
    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Spark RDD Example")
      .getOrCreate()

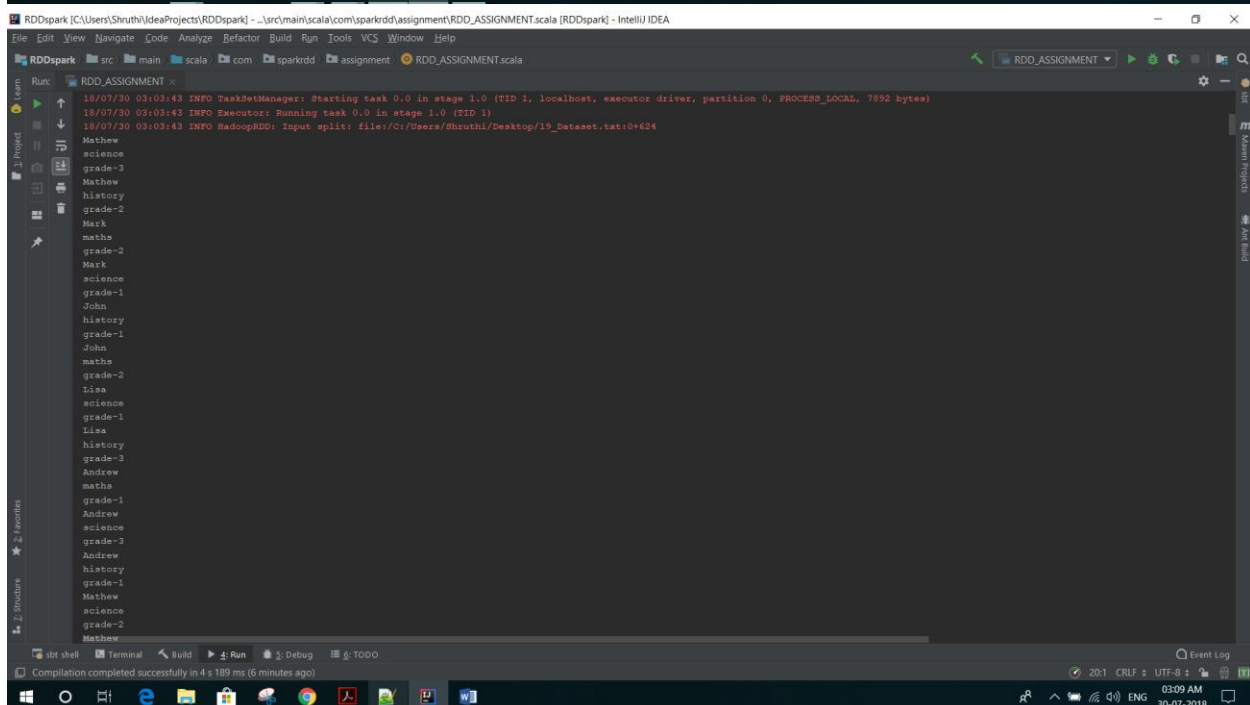
    // Task 1.1
    val newDataset = spark.sparkContext.textFile(path = "C:\\Users\\Shruthi\\Desktop\\19_Dataset.txt")
    val noOfLines = newDataset.count()
    println("no. of lines present in the file: " + noOfLines)

    // Task 1.2
    val splittedRDDs = newDataset.flatMap(f => f.split(' '))
    def isWord(s: String): Boolean = {
      val numberPattern: Regex = "[0-9]+".r
      numberPattern.findFirstMatchIn(s) match {
        case Some(_) =>
          return false
        case None =>
          return true
      }
    }
    val wordOnly = splittedRDDs.filter(f => isWord(f))
    wordOnly.foreach(s => println(s))
    val countOfWords = wordOnly.count()
  }
}

RDD_ASSIGNMENT.main(args: Array[String])
```

Run: RDD\_ASSIGNMENT x  
Process finished with exit code 0

Compilation completed successfully in 4 s 189 ms (4 minutes ago)



```
18/07/30 03:03:43 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, executor driver, partition 0, PROCESS_LOCAL, 7892 bytes)
18/07/30 03:03:43 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
18/07/30 03:03:43 INFO HadoopRDD: Input split: file:/C:/Users/Shruthi/Desktop/19_Dataset.txt:0+624
Mathew
science
grade-3
Mathew
history
grade-2
Mark
maths
grade-2
Mark
science
grade-1
John
history
grade-1
John
maths
grade-2
Lisa
science
grade-1
Lisa
history
grade-3
Andrew
maths
grade-1
Andrew
science
grade-3
Andrew
history
grade-1
Mathew
science
grade-2
Mathew
```

Run: RDD\_ASSIGNMENT x

Compilation completed successfully in 4 s 189 ms (6 minutes ago)

# RDD DEEP DIVE

2. Write a program to read a text file and print the number of words in the document.
3. We have a document where the word separator is -, instead of space. Write a spark code, to obtain the count of the total number of words present in the document.

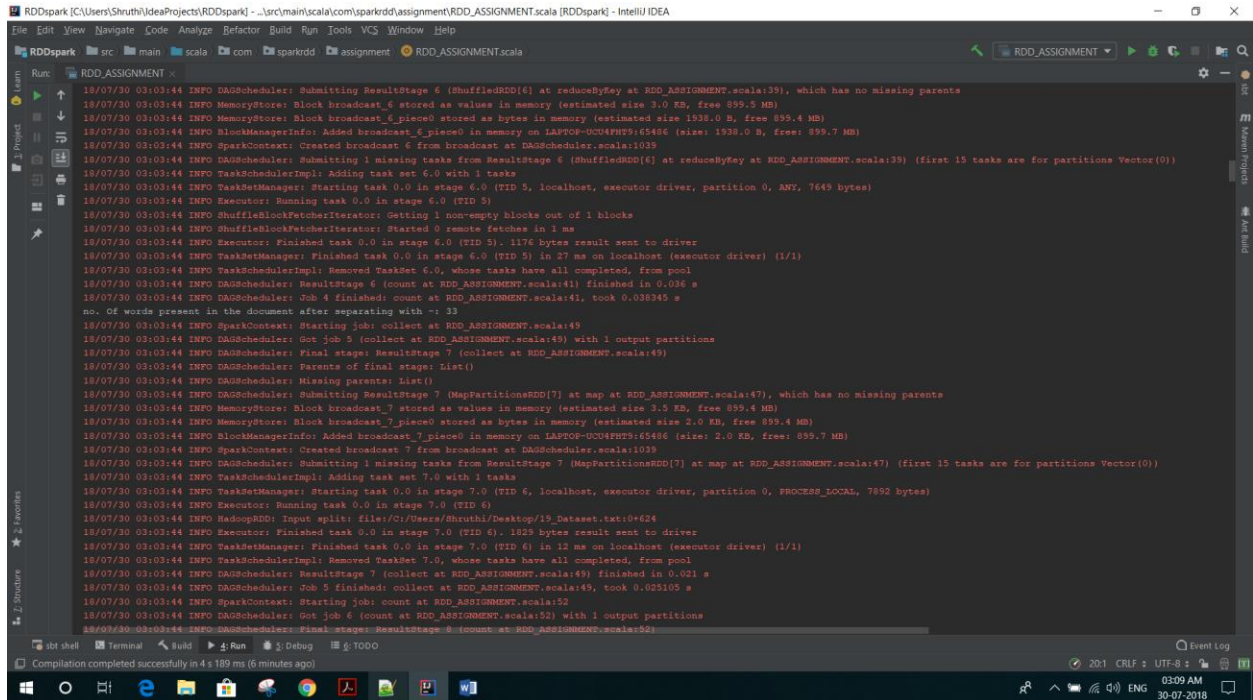
```
15 println("no. of lines present in the file: " + noOfLines)
16
17
18 //Task 1.2
19 val splittedRDDs = newDataset.flatMap(f=>f.split(' '))
20 def isWord(a:String):Boolean={
21     val numberPattern: Regex = "[0-9]+".r
22     numberPattern.findFirstMatchIn(a) match {
23         case Some(_) =>
24             return false
25         case None =>
26             return true
27     }
28 }
29 val wordOnly = splittedRDDs.filter(f=>isWord(f))
30 wordOnly.foreach(x=>println(x))
31 val countOfWords = wordOnly.count()
32 println("no. of words present in the document: " + countOfWords)
33
34 //Task 1.3
35 val splittedRDDs2 = newDataset.flatMap(f=>f.split('-'))
36 val AfterSplitting = splittedRDDs2.map(x=>(x,1)).reduceByKey(_+_ )
37 AfterSplitting.foreach(x=>println(x))
38 val countAfterSplitting = AfterSplitting.count()
39 println("no. of words present in the document after separating with -: " + countAfterSplitting)
40
41 //Task 2
42 //Problem Statement 1:
43 //1. Read the text file, and create a tuple RDD.
44 val tupleRDD = newDataset.map(x=>x.split(Regex(" ")))
45 //Printing tuple RDD
46 tupleRDD.collect().foreach(row => println(row.mkString(", ")))
47
48 //2. Find the count of total number of words present.
49 val countResult = tupleRDD.count()
50
51 RDD_ASSIGNMENT / main(args: Array[String])
52
53 Run: RDD_ASSIGNMENT x
54 Process finished with exit code 0
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
```

# RDD DEEP DIVE

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:43 INFO DAGScheduler: ResultStage 1 (foreach at RDD_ASSIGNMENT.scala:33) finished in 0.038 s
18/07/30 03:03:43 INFO DAGScheduler: Job 1 finished: foreach at RDD_ASSIGNMENT.scala:33, took 0.041901 s
18/07/30 03:03:43 INFO SparkContext: Starting job: count at RDD_ASSIGNMENT.scala:34
18/07/30 03:03:43 INFO DAGScheduler: Got job 2 (count at RDD_ASSIGNMENT.scala:34) with 1 output partitions
18/07/30 03:03:43 INFO DAGScheduler: Final stage: ResultStage 2 (count at RDD_ASSIGNMENT.scala:34)
18/07/30 03:03:43 INFO DAGScheduler: Parents of final stage: List()
18/07/30 03:03:43 INFO DAGScheduler: Missing parents: List()
18/07/30 03:03:43 INFO DAGScheduler: Submitting ResultStage 2 (MapPartitionsRDD[3] at filter at RDD_ASSIGNMENT.scala:32), which has no missing parents
18/07/30 03:03:43 INFO MemoryStore: Block broadcast_3 stored as values in memory (estimated size 3.6 KB, free 899.5 MB)
18/07/30 03:03:43 INFO MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 2.1 KB, free 899.5 MB)
18/07/30 03:03:43 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on LAPTOP-UCU4FHU1G5486 (size: 2.1 KB, free: 899.7 MB)
18/07/30 03:03:43 INFO SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:43 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 2 (MapPartitionsRDD[3] at filter at RDD_ASSIGNMENT.scala:32) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:43 INFO TaskSchedulerImpl: Adding task set 2.0 with 1 tasks
18/07/30 03:03:43 INFO TaskSetManager: Starting task 0.0 in stage 2.0 (TID 2), localhost, executor driver, partition 0, PROCESS_LOCAL, 7892 bytes)
18/07/30 03:03:43 INFO Executor: Running task 0.0 in stage 2.0 (TID 2)
18/07/30 03:03:43 INFO HadoopRDD: Input split: file:/C:/Users/Shruthi/Desktop/19_Dataset.txt:0+624
18/07/30 03:03:43 INFO Executor: Finished task 0.0 in stage 2.0 (TID 2). 789 bytes result sent to driver
18/07/30 03:03:43 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 17 ms on localhost (executor driver) (1/1)
no. Of words present in the document: 66
18/07/30 03:03:43 INFO DAGScheduler: ResultStage 2 (count at RDD_ASSIGNMENT.scala:34) finished in 0.112 s
18/07/30 03:03:43 INFO DAGScheduler: Job 2 finished: count at RDD_ASSIGNMENT.scala:34, took 0.120665 s
18/07/30 03:03:43 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
18/07/30 03:03:43 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:40
18/07/30 03:03:43 INFO DAGScheduler: Registering RDD 5 (map at RDD_ASSIGNMENT.scala:39)
18/07/30 03:03:43 INFO DAGScheduler: Got job 3 (foreach at RDD_ASSIGNMENT.scala:40) with 1 output partitions
18/07/30 03:03:43 INFO DAGScheduler: Final stage: ResultStage 4 (foreach at RDD_ASSIGNMENT.scala:40)
18/07/30 03:03:43 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 3)
18/07/30 03:03:43 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 3)
18/07/30 03:03:43 INFO DAGScheduler: Submitting ShuffleMapStage 3 (MapPartitionsRDD[5] at map at RDD_ASSIGNMENT.scala:39), which has no missing parents
18/07/30 03:03:43 INFO MemoryStore: Block broadcast_4 stored as values in memory (estimated size 4.8 KB, free 899.4 MB)
18/07/30 03:03:43 INFO MemoryStore: Block broadcast_4_piece0 stored as bytes in memory (estimated size 2.8 KB, free 899.4 MB)
18/07/30 03:03:43 INFO BlockManagerInfo: Added broadcast_4_piece0 in memory on LAPTOP-UCU4FHU1G5486 (size: 2.8 KB, free: 899.7 MB)
18/07/30 03:03:43 INFO SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:43 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 3 (MapPartitionsRDD[5] at map at RDD_ASSIGNMENT.scala:39) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:43 INFO TaskSchedulerImpl: Adding task set 3.0 with 1 tasks
18/07/30 03:03:43 INFO TaskSetManager: Starting task 0.0 in stage 3.0 (TID 3), localhost, executor driver, partition 0, PROCESS_LOCAL, 7881 bytes)
18/07/30 03:03:43 INFO Executor: Running task 0.0 in stage 3.0 (TID 3)
18/07/30 03:03:43 INFO HadoopRDD: Input split: file:/C:/Users/Shruthi/Desktop/19_Dataset.txt:0+624
18/07/30 03:03:44 INFO Executor: Finished task 0.0 in stage 3.0 (TID 3). 1156 bytes result sent to driver
Compilation completed successfully in 4 s 189 ms (6 minutes ago)
```

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:44 INFO DAGScheduler: Job 3 finished: foreach at RDD_ASSIGNMENT.scala:40, took 0.479830 s
(John,history,grade,2)
(1,34,13,1)
(Lisa,history,grade,2)
(2,55,12,1)
(1,23,16,1)
(2,97,12,1)
(1,24,12,1)
(1,67,13,1)
(1,76,13,1)
(Mark,maths,grade,2)
(Andrew,science,grade,2)
(1,35,11,1)
(3,86,13,1)
(Mathew,history,grade,2)
(2,77,11,1)
(3,44,14,1)
(John,maths,grade,2)
(3,45,12,1)
(Lisa,science,grade,2)
(2,23,13,1)
(2,96,15,1)
(Mark,science,grade,2)
(1,74,12,1)
(2,12,12,1)
(3,24,14,1)
(2,74,13,1)
(2,55,12,1)
(1,14,12,1)
(Mathew,science,grade,2)
(2,24,13,1)
(Andrew,maths,grade,2)
(Andrew,history,grade,2)
(1,92,13,1)
18/07/30 03:03:44 INFO SparkContext: Starting job: count at RDD_ASSIGNMENT.scala:41
18/07/30 03:03:44 INFO DAGScheduler: Got job 4 (count at RDD_ASSIGNMENT.scala:41) with 1 output partitions
18/07/30 03:03:44 INFO DAGScheduler: Final stage: ResultStage 6 (count at RDD_ASSIGNMENT.scala:41)
18/07/30 03:03:44 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 5)
18/07/30 03:03:44 INFO DAGScheduler: Missing parents: List()
18/07/30 03:03:44 INFO DAGScheduler: Submitting ResultStage 6 (ShuffledRDD[6] at reduceByKey at RDD_ASSIGNMENT.scala:39), which has no missing parents
Compilation completed successfully in 4 s 189 ms (6 minutes ago)
```

# RDD DEEP DIVE



The screenshot displays the IntelliJ IDEA IDE interface. The main editor window shows the file `RDD_ASSIGNMENT.scala` with the following Scala code:

```
1 RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...\src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
2
3 File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
4
5 RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
6
7 Run: RDD_ASSIGNMENT x
8
9 18/07/30 03:03:44 INFO DAGScheduler: Submitting ResultStage 6 (ShuffledRDD[4] at reduceByKey at RDD_ASSIGNMENT.scala:39), which has no missing parents
10 18/07/30 03:03:44 INFO MemoryStore: Block broadcast_6 stored as values in memory (estimated size 3.0 KB, free 899.5 MB)
11 18/07/30 03:03:44 INFO MemoryStore: Block broadcast_6_piece0 stored as bytes in memory (estimated size 1936.0 B, free 899.4 MB)
12 18/07/30 03:03:44 INFO BlockManagerInfo: Added broadcast_6_piece0 in memory on LAPTOP-UC04PHUTY (6486 (size: 1936.0 B, free: 899.7 MB)
13 18/07/30 03:03:44 INFO SparkContext: Created broadcast 6 from broadcast at DAGScheduler.scala:1039
14 18/07/30 03:03:44 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 6 (ShuffledRDD[4] at reduceByKey at RDD_ASSIGNMENT.scala:39) (first 15 tasks are for partitions Vector(0))
15 18/07/30 03:03:44 INFO TaskSchedulerImpl: Adding task set 6.0 with 1 tasks
16 18/07/30 03:03:44 INFO TaskSetManager: Starting task 0.0 in stage 6.0 (TID 5, localhost, executor driver, partition 0, ANY, 7649 bytes)
17 18/07/30 03:03:44 INFO Executor: Running task 0.0 in stage 6.0 (TID 5)
18 18/07/30 03:03:44 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
19 18/07/30 03:03:44 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
20 18/07/30 03:03:44 INFO Executor: Finished task 0.0 in stage 6.0 (TID 5): 1174 bytes result sent to driver
21 18/07/30 03:03:44 INFO TaskSetManager: Finished task 0.0 in stage 6.0 (TID 5) in 27 ms on localhost (executor driver) (1/1)
22 18/07/30 03:03:44 INFO TaskSchedulerImpl: Removed TaskSet 6.0, whose tasks have all completed, from pool
23 18/07/30 03:03:44 INFO DAGScheduler: ResultStage 6 (count at RDD_ASSIGNMENT.scala:41) finished in 0.036 s
24 18/07/30 03:03:44 INFO DAGScheduler: Job 4 finished: count at RDD_ASSIGNMENT.scala:41, took 0.038345 s
25 no. of words present in the document after separating with -r: 33
26 18/07/30 03:03:44 INFO SparkContext: Starting job: collect at RDD_ASSIGNMENT.scala:49
27 18/07/30 03:03:44 INFO DAGScheduler: Got job 5 (collect at RDD_ASSIGNMENT.scala:49) with 1 output partitions
28 18/07/30 03:03:44 INFO DAGScheduler: Final stage: ResultStage 7 (collect at RDD_ASSIGNMENT.scala:49)
29 18/07/30 03:03:44 INFO DAGScheduler: Parents of final stage: List()
30 18/07/30 03:03:44 INFO DAGScheduler: Missing parents: List()
31 18/07/30 03:03:44 INFO DAGScheduler: Submitting ResultStage 7 (MapPartitionsRDD[7] at map at RDD_ASSIGNMENT.scala:47), which has no missing parents
32 18/07/30 03:03:44 INFO MemoryStore: Block broadcast_7 stored as values in memory (estimated size 3.5 KB, free 899.4 MB)
33 18/07/30 03:03:44 INFO MemoryStore: Block broadcast_7_piece0 stored as bytes in memory (estimated size 2.0 KB, free 899.4 MB)
34 18/07/30 03:03:44 INFO BlockManagerInfo: Added broadcast_7_piece0 in memory on LAPTOP-UC04PHUTY (6486 (size: 2.0 KB, free: 899.7 MB)
35 18/07/30 03:03:44 INFO SparkContext: Created broadcast 7 from broadcast at DAGScheduler.scala:1039
36 18/07/30 03:03:44 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 7 (MapPartitionsRDD[7] at map at RDD_ASSIGNMENT.scala:47) (first 15 tasks are for partitions Vector(0))
37 18/07/30 03:03:44 INFO TaskSchedulerImpl: Adding task set 7.0 with 1 tasks
38 18/07/30 03:03:44 INFO TaskSetManager: Starting task 0.0 in stage 7.0 (TID 6, localhost, executor driver, partition 0, PROCESS_LOCAL, 7892 bytes)
39 18/07/30 03:03:44 INFO Executor: Running task 0.0 in stage 7.0 (TID 6)
40 18/07/30 03:03:44 INFO HadoopRDD: Input split: file:/C:/Users/Shruthi/Desktop/19_Dataset.txt:0+624
41 18/07/30 03:03:44 INFO Executor: Finished task 0.0 in stage 7.0 (TID 6): 1629 bytes result sent to driver
42 18/07/30 03:03:44 INFO TaskSetManager: Finished task 0.0 in stage 7.0 (TID 6) in 12 ms on localhost (executor driver) (1/1)
43 18/07/30 03:03:44 INFO TaskSchedulerImpl: Removed TaskSet 7.0, whose tasks have all completed, from pool
44 18/07/30 03:03:44 INFO DAGScheduler: ResultStage 7 (collect at RDD_ASSIGNMENT.scala:49) finished in 0.021 s
45 18/07/30 03:03:44 INFO DAGScheduler: Job 5 finished: collect at RDD_ASSIGNMENT.scala:49, took 0.025105 s
46 18/07/30 03:03:44 INFO SparkContext: Starting job: count at RDD_ASSIGNMENT.scala:52
47 18/07/30 03:03:44 INFO DAGScheduler: Got job 6 (count at RDD_ASSIGNMENT.scala:52) with 1 output partitions
48 18/07/30 03:03:44 INFO DAGScheduler: Final stage: ResultStage 8 (count at RDD_ASSIGNMENT.scala:52)
```

The bottom status bar indicates: `Compilation completed successfully in 4 s 189 ms (6 minutes ago)`. The system tray shows the date and time as 03:09 AM, 30-07-2018.

# RDD DEEP DIVE

## Task 2

### Problem Statement 1:

1. Read the text file, and create a tupled rdd.
2. Find the count of total number of rows present.
3. What is the distinct number of subjects present in the entire school
4. What is the count of the number of students in the school, whose name is Mathew and marks is 55

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - \src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
//Task 2
//Problem Statement 1:
//1. Read the text file, and create a tupled rdd.
val tupledRDD = newDataset.map(x=>x.split(" ")).map(x=>(x(0),x(1)))
//printing tupled RDD
tupledRDD.collect().foreach(row => println(row.mkString(", ")))

//2. Find the count of total number of rows present.
val countResult = tupledRDD.count()
println("No. Of rows present in tupled RDD -> " + countResult)

//3. What is the distinct number of subjects present in the entire school
val subjectOnly = tupledRDD.map(item=>item(1))
//get distinct subjects
val distinctSubjects = subjectOnly.distinct()
distinctSubjects.foreach(s=>println(s))

//4. What is the count of the number of students in the school, whose name is Mathew and marks is 55
def getStudent(name:String,marks:Int) : Boolean={
  if (name.equalsIgnoreCase("Mathew") && marks==55)
    return true
  else
    return false
}

val students = tupledRDD.filter(x=>getStudent(x(0),x(1).toInt))
students.collect().foreach(row => println(row.mkString(", ")))
val countStudents = students.count()
println("Number of students in the school, whose name is Mathew and marks is 55 -> " + countStudents)

//Problem Statement 2:
//1. What is the count of students per grade in the school?
val stdWithGrades = tupledRDD.map(x=>(x(2),1))

RDD_ASSIGNMENT main(args: Array[String])
Run: RDD_ASSIGNMENT
Process finished with exit code 0
Compilation completed successfully in 4 s 189 ms (5 minutes ago)
```

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - \src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT
18/07/30 03:03:44 INFO DAGScheduler: Job 5 finished: collect at RDD_ASSIGNMENT.scala:49, took 0.025105 s
18/07/30 03:03:44 INFO SparkContext: Starting job: count at RDD_ASSIGNMENT.scala:52
18/07/30 03:03:44 INFO DAGScheduler: Got job 6 (count at RDD_ASSIGNMENT.scala:52) with 1 output partitions
18/07/30 03:03:44 INFO DAGScheduler: Final stage: ResultStage 0 (count at RDD_ASSIGNMENT.scala:52)
18/07/30 03:03:44 INFO DAGScheduler: Parents of final stage: List()
18/07/30 03:03:44 INFO DAGScheduler: Missing parents: List()
18/07/30 03:03:44 INFO DAGScheduler: Submitting ResultStage 0 (MapPartitionsRDD[7] at map at RDD_ASSIGNMENT.scala:47), which has no missing parents
Mathew,science,grade-3,45,12
Mathew,history,grade-2,55,13
Mark,maths,grade-2,23,13
Mark,science,grade-1,76,13
John,history,grade-1,14,12
John,maths,grade-2,74,13
Lisa,science,grade-1,24,12
Lisa,history,grade-3,96,13
Andrew,maths,grade-1,34,13
Andrew,science,grade-3,26,14
Andrew,history,grade-1,74,12
Mathew,science,grade-2,55,12
Mathew,history,grade-2,87,12
Mark,maths,grade-1,92,13
Mark,science,grade-2,12,12
John,history,grade-1,67,13
John,maths,grade-1,35,11
Lisa,science,grade-2,24,13
Lisa,history,grade-2,98,15
Andrew,maths,grade-1,23,16
Andrew,science,grade-2,44,14
Andrew,history,grade-2,77,11
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size 3.3 KB, free 895.4 MB)
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 2038.0 B, free 895.4 MB)
18/07/30 03:03:44 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on LAPTOP-UCU4FHT9:65406 (size: 2038.0 B, free: 895.7 MB)
18/07/30 03:03:44 INFO SparkContext: Created broadcast 0 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:44 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 0 (MapPartitionsRDD[7] at map at RDD_ASSIGNMENT.scala:47) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:44 INFO TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
18/07/30 03:03:44 INFO TaskSchedulerImpl: Starting task 0.0 in stage 0.0 (TID 7, localhost, executor driver, partition 0, PROCESS_LOCAL, 7892 bytes)
18/07/30 03:03:44 INFO Executor: Running task 0.0 in stage 0.0 (TID 7)
18/07/30 03:03:44 INFO HadoopRDD: Input split: file:/C:/Users/Shruthi/Desktop/19_Dataset.txt:0+624
18/07/30 03:03:44 INFO Executor: Finished task 0.0 in stage 0.0 (TID 7). 746 bytes result sent to driver
18/07/30 03:03:44 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 7) in 7 ms on localhost (executor driver: 1/1)
Compilation completed successfully in 4 s 189 ms (6 minutes ago)
```



# RDD DEEP DIVE

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
Andrew.science.grade=3,44,14
Andrew.history.grade=2,77,11
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_8 stored as values in memory (estimated size 3.3 KB, free 899.4 MB)
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_9_piece0 stored as bytes in memory (estimated size 2038.0 B, free 899.4 MB)
18/07/30 03:03:44 INFO BlockManagerInfo: Added broadcast_9_piece0 in memory on LAPTOP-UCM4PH99:65486 (size: 2038.0 B, free: 899.7 MB)
18/07/30 03:03:44 INFO SparkContext: Created broadcast 9 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:44 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 8 (MapPartitionsRDD[7] at map at RDD_ASSIGNMENT.scala:47) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:44 INFO TaskSchedulerImpl: Adding task set 8.0 with 1 tasks
18/07/30 03:03:44 INFO TaskSetManager: Starting task 0.0 in stage 8.0 (TID 7, localhost, executor driver, partition 0, PROCESS_LOCAL, 7892 bytes)
18/07/30 03:03:44 INFO Executor: Running task 0.0 in stage 8.0 (TID 7)
18/07/30 03:03:44 INFO HadoopRDD: Input split: file:/C:/Users/Shruthi/Desktop/19_Dataset.txt:0+624
18/07/30 03:03:44 INFO Executor: Finished task 0.0 in stage 8.0 (TID 7), 744 bytes result sent to driver
18/07/30 03:03:44 INFO TaskSetManager: Finished task 0.0 in stage 8.0 (TID 7) in 7 ms on localhost (executor driver) (1/1)
18/07/30 03:03:44 INFO TaskSchedulerImpl: Removed TaskSet 8.0, whose tasks have all completed, from pool
18/07/30 03:03:44 INFO DAGScheduler: ResultStage 8 (count at RDD_ASSIGNMENT.scala:52) finished in 0.014 s
18/07/30 03:03:44 INFO DAGScheduler: Job 6 finished: count at RDD_ASSIGNMENT.scala:52, took 0.016122 s
no. Of rows present in tupled RDD => 22
18/07/30 03:03:44 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:55
18/07/30 03:03:44 INFO DAGScheduler: Submitting RDD 9 (distinct at RDD_ASSIGNMENT.scala:55)
18/07/30 03:03:44 INFO DAGScheduler: Got job 7 (foreach at RDD_ASSIGNMENT.scala:55) with 1 output partitions
18/07/30 03:03:44 INFO DAGScheduler: Final stage: ResultStage 10 (foreach at RDD_ASSIGNMENT.scala:55)
18/07/30 03:03:44 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 9)
18/07/30 03:03:44 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 9)
18/07/30 03:03:44 INFO DAGScheduler: Submitting ShuffleMapStage 9 (MapPartitionsRDD[9] at distinct at RDD_ASSIGNMENT.scala:58), which has no missing parents
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_9 stored as values in memory (estimated size 4.9 KB, free 899.4 MB)
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_9_piece0 stored as bytes in memory (estimated size 2.8 KB, free 899.4 MB)
18/07/30 03:03:44 INFO BlockManagerInfo: Added broadcast_9_piece0 in memory on LAPTOP-UCM4PH99:65486 (size: 2.8 KB, free: 899.7 MB)
18/07/30 03:03:44 INFO SparkContext: Created broadcast 9 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:44 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 9 (MapPartitionsRDD[9] at distinct at RDD_ASSIGNMENT.scala:58) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:44 INFO TaskSchedulerImpl: Adding task set 9.0 with 1 tasks
18/07/30 03:03:44 INFO TaskSetManager: Starting task 0.0 in stage 9.0 (TID 8, localhost, executor driver, partition 0, PROCESS_LOCAL, 7881 bytes)
18/07/30 03:03:44 INFO Executor: Running task 0.0 in stage 9.0 (TID 8)
18/07/30 03:03:44 INFO HadoopRDD: Input split: file:/C:/Users/Shruthi/Desktop/19_Dataset.txt:0+624
18/07/30 03:03:44 INFO Executor: Finished task 0.0 in stage 9.0 (TID 8), 1070 bytes result sent to driver
18/07/30 03:03:44 INFO TaskSetManager: Finished task 0.0 in stage 9.0 (TID 8) in 26 ms on localhost (executor driver) (1/1)
18/07/30 03:03:44 INFO TaskSchedulerImpl: Removed TaskSet 9.0, whose tasks have all completed, from pool
18/07/30 03:03:44 INFO DAGScheduler: ShuffleMapStage 9 (distinct at RDD_ASSIGNMENT.scala:58) finished in 0.034 s
18/07/30 03:03:44 INFO DAGScheduler: looking for newly runnable stages
18/07/30 03:03:44 INFO DAGScheduler: running: Set()
18/07/30 03:03:44 INFO DAGScheduler: waiting: Set(ResultStage 10)
Compilation completed successfully in 4 s 189 ms (6 minutes ago)
```

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:44 INFO DAGScheduler: waiting: Set(ResultStage 10)
18/07/30 03:03:44 INFO DAGScheduler: failed: Set()
18/07/30 03:03:44 INFO DAGScheduler: Submitting ResultStage 10 (MapPartitionsRDD[11] at distinct at RDD_ASSIGNMENT.scala:58), which has no missing parents
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_10 stored as values in memory (estimated size 3.9 KB, free 899.4 MB)
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_10_piece0 stored as bytes in memory (estimated size 2.3 KB, free 899.4 MB)
18/07/30 03:03:44 INFO BlockManagerInfo: Added broadcast_10_piece0 in memory on LAPTOP-UCM4PH99:65486 (size: 2.3 KB, free: 899.7 MB)
18/07/30 03:03:44 INFO SparkContext: Created broadcast 10 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:44 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 10 (MapPartitionsRDD[11] at distinct at RDD_ASSIGNMENT.scala:58) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:44 INFO TaskSchedulerImpl: Adding task set 10.0 with 1 tasks
18/07/30 03:03:44 INFO TaskSetManager: Starting task 0.0 in stage 10.0 (TID 9, localhost, executor driver, partition 0, ANY, 7649 bytes)
18/07/30 03:03:44 INFO Executor: Running task 0.0 in stage 10.0 (TID 9)
18/07/30 03:03:44 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:44 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/30 03:03:44 INFO Executor: Finished task 0.0 in stage 10.0 (TID 9), 1095 bytes result sent to driver
18/07/30 03:03:44 INFO TaskSetManager: Finished task 0.0 in stage 10.0 (TID 9) in 17 ms on localhost (executor driver) (1/1)
18/07/30 03:03:44 INFO TaskSchedulerImpl: Removed TaskSet 10.0, whose tasks have all completed, from pool
18/07/30 03:03:44 INFO DAGScheduler: ResultStage 10 (foreach at RDD_ASSIGNMENT.scala:55) finished in 0.034 s
maths
history
18/07/30 03:03:44 INFO DAGScheduler: Job 7 finished: foreach at RDD_ASSIGNMENT.scala:59, took 0.074940 s
science
18/07/30 03:03:44 INFO SparkContext: Starting job: collect at RDD_ASSIGNMENT.scala:70
18/07/30 03:03:44 INFO DAGScheduler: Got job 8 (collect at RDD_ASSIGNMENT.scala:70) with 1 output partitions
18/07/30 03:03:44 INFO DAGScheduler: Final stage: ResultStage 11 (collect at RDD_ASSIGNMENT.scala:70)
18/07/30 03:03:44 INFO DAGScheduler: Parents of final stage: List()
18/07/30 03:03:44 INFO DAGScheduler: Missing parents: List()
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_11 stored as values in memory (estimated size 3.7 KB, free 899.4 MB)
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_11_piece0 stored as bytes in memory (estimated size 2.1 KB, free 899.4 MB)
18/07/30 03:03:44 INFO BlockManagerInfo: Added broadcast_11_piece0 in memory on LAPTOP-UCM4PH99:65486 (size: 2.1 KB, free: 899.7 MB)
18/07/30 03:03:44 INFO SparkContext: Created broadcast 11 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:44 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 11 (MapPartitionsRDD[12] at filter at RDD_ASSIGNMENT.scala:69) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:44 INFO TaskSchedulerImpl: Adding task set 11.0 with 1 tasks
18/07/30 03:03:44 INFO TaskSetManager: Starting task 0.0 in stage 11.0 (TID 10, localhost, executor driver, partition 0, PROCESS_LOCAL, 7892 bytes)
18/07/30 03:03:44 INFO Executor: Running task 0.0 in stage 11.0 (TID 10)
18/07/30 03:03:44 INFO HadoopRDD: Input split: file:/C:/Users/Shruthi/Desktop/19_Dataset.txt:0+624
18/07/30 03:03:44 INFO Executor: Finished task 0.0 in stage 11.0 (TID 10), 880 bytes result sent to driver
18/07/30 03:03:44 INFO TaskSetManager: Finished task 0.0 in stage 11.0 (TID 10) in 8 ms on localhost (executor driver) (1/1)
18/07/30 03:03:44 INFO TaskSchedulerImpl: Removed TaskSet 11.0, whose tasks have all completed, from pool
18/07/30 03:03:44 INFO DAGScheduler: ResultStage 11 (collect at RDD_ASSIGNMENT.scala:70) finished in 0.046 s
Compilation completed successfully in 4 s 189 ms (7 minutes ago)
```

# RDD DEEP DIVE

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:44 INFO SparkContext: Starting job: collect at RDD_ASSIGNMENT.scala:70
18/07/30 03:03:44 INFO DAGScheduler: Got job 8 (collect at RDD_ASSIGNMENT.scala:70) with 1 output partitions
18/07/30 03:03:44 INFO DAGScheduler: Final stage: ResultStage 11 (collect at RDD_ASSIGNMENT.scala:70)
18/07/30 03:03:44 INFO DAGScheduler: Parents of final stage: List()
18/07/30 03:03:44 INFO DAGScheduler: Submitting ResultStage 11 (MapPartitionsRDD[12] at filter at RDD_ASSIGNMENT.scala:69), which has no missing parents
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_11_piece0 stored as values in memory (estimated size 3.7 KB, free 899.4 MB)
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_11_piece0 stored as bytes in memory (estimated size 2.1 KB, free 899.4 MB)
18/07/30 03:03:44 INFO BlockManagerInfo: Added broadcast_11_piece0 in memory on LAPTOP-UCD4PHT9:65486 (size: 2.1 KB, free: 899.7 MB)
18/07/30 03:03:44 INFO SparkContext: Created broadcast 11 from broadcast at DAGScheduler.scala:1035
18/07/30 03:03:44 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 11 (MapPartitionsRDD[12] at filter at RDD_ASSIGNMENT.scala:69) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:44 INFO TaskSchedulerImpl: Adding task set 11.0 with 1 tasks
18/07/30 03:03:44 INFO TaskSetManager: Starting task 0.0 in stage 11.0 (TID 10, localhost, executor driver, partition 0, PROCESS_LOCAL, 7892 bytes)
18/07/30 03:03:44 INFO Executor: Running task 0.0 in stage 11.0 (TID 10)
18/07/30 03:03:44 INFO HadoopRDD: Input split: file:/C:/Users/Shruthi/Desktop/19_Dataset.txt:0+624
18/07/30 03:03:44 INFO Executor: Finished task 0.0 in stage 11.0 (TID 10). 880 bytes result sent to driver
18/07/30 03:03:44 INFO TaskSetManager: Finished task 0.0 in stage 11.0 (TID 10) in 8 ms on localhost (executor driver) (1/1)
18/07/30 03:03:44 INFO TaskSchedulerImpl: Removed TaskSet 11.0, whose tasks have all completed, from pool
18/07/30 03:03:44 INFO DAGScheduler: ResultStage 11 (collect at RDD_ASSIGNMENT.scala:70) finished in 0.016 s
18/07/30 03:03:44 INFO DAGScheduler: Job 8 finished: collect at RDD_ASSIGNMENT.scala:70, took 0.020407 s
18/07/30 03:03:44 INFO SparkContext: Starting job: count at RDD_ASSIGNMENT.scala:71
Mathew.history.grade-2,55,13
18/07/30 03:03:44 INFO DAGScheduler: Got job 9 (count at RDD_ASSIGNMENT.scala:71) with 1 output partitions
Mathew.science.grade-2,55,12
18/07/30 03:03:44 INFO DAGScheduler: Final stage: ResultStage 12 (count at RDD_ASSIGNMENT.scala:71)
18/07/30 03:03:44 INFO DAGScheduler: Parents of final stage: List()
18/07/30 03:03:44 INFO DAGScheduler: Submitting ResultStage 12 (MapPartitionsRDD[12] at filter at RDD_ASSIGNMENT.scala:69), which has no missing parents
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_12_piece0 stored as values in memory (estimated size 3.5 KB, free 899.4 MB)
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_12_piece0 stored as bytes in memory (estimated size 2.1 KB, free 899.4 MB)
18/07/30 03:03:44 INFO BlockManagerInfo: Added broadcast_12_piece0 in memory on LAPTOP-UCD4PHT9:65486 (size: 2.1 KB, free: 899.7 MB)
18/07/30 03:03:44 INFO SparkContext: Created broadcast 12 from broadcast at DAGScheduler.scala:1035
18/07/30 03:03:44 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 12 (MapPartitionsRDD[12] at filter at RDD_ASSIGNMENT.scala:69) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:44 INFO TaskSchedulerImpl: Adding task set 12.0 with 1 tasks
18/07/30 03:03:44 INFO TaskSetManager: Starting task 0.0 in stage 12.0 (TID 11, localhost, executor driver, partition 0, PROCESS_LOCAL, 7892 bytes)
18/07/30 03:03:44 INFO Executor: Running task 0.0 in stage 12.0 (TID 11)
18/07/30 03:03:44 INFO HadoopRDD: Input split: file:/C:/Users/Shruthi/Desktop/19_Dataset.txt:0+624
18/07/30 03:03:44 INFO Executor: Finished task 0.0 in stage 12.0 (TID 11). 746 bytes result sent to driver
18/07/30 03:03:44 INFO TaskSetManager: Finished task 0.0 in stage 12.0 (TID 11) in 9 ms on localhost (executor driver) (1/1)
18/07/30 03:03:44 INFO TaskSchedulerImpl: Removed TaskSet 12.0, whose tasks have all completed, from pool
Compilation completed successfully in 4 s 189 ms (7 minutes ago)
201 CRLF UTF-8 03:10 AM 30-07-2018
```

# RDD DEEP DIVE

## Problem Statement 2:

1. What is the count of students per grade in the school?
2. Find the average of each student (Note - Mathew is grade-1, is different from Mathew in some other grade!)
3. What is the average score of students in each subject across all grades?
4. What is the average score of students in each subject per grade?
5. For all students in grade-2, how many have average score greater than 50?

```
RDDspark [C:\Users\Shruti\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkdd assignment RDD_ASSIGNMENT.scala
build.sbt x RDD_ASSIGNMENT.scala
//Problem Statement 2:
//1. What is the count of students per grade in the school?
val stdsWithGrades = tupleRDD.map(x=>(x(2),1))
val group = stdsWithGrades.groupByKey()
group.foreach(x=>println("Students in Grade:" + x))

//Variation : Getting the distinct student counts per grade depending upon his name
val distinctStd = tupleRDD.map(x=>(x(2),x(0)))
val group1 = distinctStd.distinct.groupByKey()
group1.foreach(x=>println(x))

//2. Find the average of each student (Note - Mathew is grade-1, is different from Mathew in
//some other grade!)
val student1 = tupleRDD.map(x=>((x(2),x(0)),x(3).toInt))
val sum1 = student1.distinct.groupByKey().mapValues(x=>x.sum)
sum1.foreach(x=>println(x))
val student2 = tupleRDD.map(x=>((x(2),x(0)),1))
//Sum up the counter to get the subjects count
val count1 = student2.groupByKey().mapValues(x=>x.sum)
count1.foreach(x=>println(x))
//We have to join these two RDDs to get the Sum and count in one RDD so that we can get the Average of each student
val joinedRDD = sum1.join(count1)
joinedRDD.foreach(x=>println(x))
//calculate average
val averageOfEachStd = joinedRDD.map(x=>((x._1._1,x._2._1/x._2._2)))
averageOfEachStd.foreach(x=>println("Average Of Student: " + x))

//3. What is the average score of students in each subject across all grades?
val subjectMarks = tupleRDD.map(x=>(x(1),x(3).toInt))
val sum3 = subjectMarks.groupByKey().mapValues(x=>x.sum)
sum3.foreach(x=>println(x))
val noOfStudents = tupleRDD.map(x=>(x(1),1)).groupByKey().mapValues(x=>x.sum)
noOfStudents.foreach(x=>println(x))

RDD_ASSIGNMENT main(args: Array[String])
Run: RDD_ASSIGNMENT x
Process finished with exit code 0

sbt shell Terminal Build Run Debug TOODO
Compilation completed successfully in 4 s 189 ms (5 minutes ago)

RDDspark [C:\Users\Shruti\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkdd assignment RDD_ASSIGNMENT.scala
build.sbt x RDD_ASSIGNMENT.scala
//3. What is the average score of students in each subject across all grades?
val subjectMarks = tupleRDD.map(x=>(x(1),x(3).toInt))
val sum3 = subjectMarks.groupByKey().mapValues(x=>x.sum)
sum3.foreach(x=>println(x))
val noOfStudents = tupleRDD.map(x=>(x(1),1)).groupByKey().mapValues(x=>x.sum)
noOfStudents.foreach(x=>println(x))
//Join these RDD and calculate the average
val avg2 = sum3.join(noOfStudents).map(x=>((x._1._1,x._2._1/x._2._2)))
avg2.foreach(x=>println("Average for Subject:" + x._1 + "=" + x._2))

//4. What is the average score of students in each subject per grade?
val subMarksPerGrade = tupleRDD.map(x=>((x(1),x(2)),x(3).toInt))
val sum4 = subMarksPerGrade.groupByKey().mapValues(x=>x.sum)
sum4.foreach(x=>println(x))
//Calculate count of students
val noOfStudents1 = tupleRDD.map(x=>((x(1),x(2)),1)).groupByKey().mapValues(x=>x.sum)
noOfStudents1.foreach(x=>println(x))
//Join these RDD and calculate the average
val avg3 = sum4.join(noOfStudents1).map(x=>((x._1._1,x._2._1/x._2._2)))
avg3.foreach(x=>println("Average for : " + x._1 + "=" + x._2))

//5. For all students in grade-2, how many have average score greater than 50?
val grade2Students = averageOfEachStd.filter(x=>x._1._1.equalsIgnoreCase("grade-2") && x._2 > 50)
grade2Students.foreach(x=>println(x))

//Problem Statement 3:
//1. Average score per student_name across all grades is same as average score per student_name per grade
//Hint - Use Interaction Property
val sumPerStdName = tupleRDD.map(x=>(x(0),x(3).toInt)).groupByKey().mapValues(y=>y.sum)
sumPerStdName.foreach(x=>println(x))
val countOfStdName = tupleRDD.map(x=>(x(0),1)).groupByKey().mapValues(y=>y.sum)

RDD_ASSIGNMENT main(args: Array[String])
Run: RDD_ASSIGNMENT x
Process finished with exit code 0

sbt shell Terminal Build Run Debug TOODO
Compilation completed successfully in 4 s 189 ms (5 minutes ago)
```



# RDD DEEP DIVE

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - \src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:44 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 14 (ShuffledRDD[15] at countByKey at RDD_ASSIGNMENT.scala:77) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:44 INFO TaskSchedulerImpl: Adding task set 14.0 with 1 tasks
18/07/30 03:03:44 INFO TaskSetManager: Starting task 0.0 in stage 14.0 (TID 13, localhost, executor driver, partition 0, ANY, 7649 bytes)
18/07/30 03:03:44 INFO Executor: Running task 0.0 in stage 14.0 (TID 13)
18/07/30 03:03:44 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:44 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
18/07/30 03:03:44 INFO Executor: Finished task 0.0 in stage 14.0 (TID 13). 1338 bytes result sent to driver
18/07/30 03:03:44 INFO TaskSetManager: Finished task 0.0 in stage 14.0 (TID 13) in 16 ms on localhost (executor driver) (1/1)
18/07/30 03:03:44 INFO TaskSchedulerImpl: Removed TaskSet 14.0, whose tasks have all completed, from pool
18/07/30 03:03:44 INFO DAGScheduler: ResultStage 14 (countByKey at RDD_ASSIGNMENT.scala:77) finished in 0.022 s
18/07/30 03:03:44 INFO DAGScheduler: Job 10 finished: countByKey at RDD_ASSIGNMENT.scala:77, took 0.052332 s
Students in Grade: (grade=1,9)
Students in Grade: (grade=2,9)
18/07/30 03:03:44 INFO SparkContext: Starting job: countByKey at RDD_ASSIGNMENT.scala:82
18/07/30 03:03:44 INFO DAGScheduler: Registering RDD 17 (distinct at RDD_ASSIGNMENT.scala:82)
18/07/30 03:03:44 INFO DAGScheduler: Registering RDD 20 (countByKey at RDD_ASSIGNMENT.scala:82)
18/07/30 03:03:44 INFO DAGScheduler: Got job 11 (countByKey at RDD_ASSIGNMENT.scala:82) with 1 output partitions
18/07/30 03:03:44 INFO DAGScheduler: Final stage: ResultStage 17 (countByKey at RDD_ASSIGNMENT.scala:82)
18/07/30 03:03:44 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 16)
18/07/30 03:03:44 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 16)
18/07/30 03:03:44 INFO DAGScheduler: Submitting ShuffleMapStage 15 (MapPartitionsRDD[17] at distinct at RDD_ASSIGNMENT.scala:82), which has no missing parents
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 215
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 338
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 329
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 193
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 184
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 182
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 188
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 142
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 217
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 262
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 199
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 246
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 155
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 250
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 227
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 163
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 322
18/07/30 03:03:44 INFO ContextCleaner: Cleaned accumulator 194
Compilation completed successfully in 4 s 189 ms (7 minutes ago)
```

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - \src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:44 INFO DAGScheduler: Submitting ResultStage 17 (ShuffledRDD[21] at countByKey at RDD_ASSIGNMENT.scala:82), which has no missing parents
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_17 stored as values in memory (estimated size 3.2 KB, free 859.4 MB)
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_17_piece0 stored as bytes in memory (estimated size 1946.0 B, free 859.4 MB)
18/07/30 03:03:44 INFO BlockManagerInfo: Added broadcast_17_piece0 in memory on LAPTOP-UCU4PH79:85486 (size: 1946.0 B, free: 859.7 MB)
18/07/30 03:03:44 INFO SparkContext: Created broadcast 17 from broadcast at DAGScheduler.scala:1035
18/07/30 03:03:44 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 17 (ShuffledRDD[21] at countByKey at RDD_ASSIGNMENT.scala:82) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:44 INFO TaskSchedulerImpl: Adding task set 17.0 with 1 tasks
18/07/30 03:03:44 INFO TaskSetManager: Starting task 0.0 in stage 17.0 (TID 16, localhost, executor driver, partition 0, ANY, 7649 bytes)
18/07/30 03:03:44 INFO Executor: Running task 0.0 in stage 17.0 (TID 16)
18/07/30 03:03:44 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:44 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/30 03:03:44 INFO Executor: Finished task 0.0 in stage 17.0 (TID 16). 1307 bytes result sent to driver
18/07/30 03:03:44 INFO TaskSetManager: Finished task 0.0 in stage 17.0 (TID 16) in 15 ms on localhost (executor driver) (1/1)
18/07/30 03:03:44 INFO TaskSchedulerImpl: Removed TaskSet 17.0, whose tasks have all completed, from pool
(grade=3,3)
18/07/30 03:03:44 INFO DAGScheduler: ResultStage 17 (countByKey at RDD_ASSIGNMENT.scala:82) finished in 0.023 s
18/07/30 03:03:44 INFO DAGScheduler: Job 11 finished: countByKey at RDD_ASSIGNMENT.scala:82, took 0.164187 s
(grade=1,4)
(grade=2,5)
18/07/30 03:03:44 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:89
18/07/30 03:03:44 INFO DAGScheduler: Registering RDD 23 (distinct at RDD_ASSIGNMENT.scala:88)
18/07/30 03:03:44 INFO DAGScheduler: Registering RDD 25 (distinct at RDD_ASSIGNMENT.scala:88)
18/07/30 03:03:44 INFO DAGScheduler: Got job 12 (foreach at RDD_ASSIGNMENT.scala:89) with 1 output partitions
18/07/30 03:03:44 INFO DAGScheduler: Final stage: ResultStage 20 (foreach at RDD_ASSIGNMENT.scala:89)
18/07/30 03:03:44 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 19)
18/07/30 03:03:44 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 19)
18/07/30 03:03:44 INFO SparkContext: Submitting ShuffleMapStage 18 (MapPartitionsRDD[23] at distinct at RDD_ASSIGNMENT.scala:88), which has no missing parents
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_18 stored as values in memory (estimated size 4.5 KB, free 859.4 MB)
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_18_piece0 stored as bytes in memory (estimated size 2.6 KB, free 859.4 MB)
18/07/30 03:03:44 INFO BlockManagerInfo: Added broadcast_18_piece0 in memory on LAPTOP-UCU4PH79:85486 (size: 2.6 KB, free: 859.7 MB)
18/07/30 03:03:44 INFO SparkContext: Created broadcast 18 from broadcast at DAGScheduler.scala:1035
18/07/30 03:03:44 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 18 (MapPartitionsRDD[23] at distinct at RDD_ASSIGNMENT.scala:88) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:44 INFO TaskSchedulerImpl: Adding task set 18.0 with 1 tasks
18/07/30 03:03:44 INFO TaskSetManager: Starting task 0.0 in stage 18.0 (TID 17, localhost, executor driver, partition 0, PROCESS_LOCAL, 7881 bytes)
18/07/30 03:03:44 INFO Executor: Running task 0.0 in stage 18.0 (TID 17)
18/07/30 03:03:44 INFO HadoopRDD: Input split: file:/C:/Users/Shruthi/Desktop/19_Dataset.txt:0+624
18/07/30 03:03:44 INFO Executor: Finished task 0.0 in stage 18.0 (TID 17). 1070 bytes result sent to driver
18/07/30 03:03:44 INFO TaskSetManager: Finished task 0.0 in stage 18.0 (TID 17) in 21 ms on localhost (executor driver) (1/1)
18/07/30 03:03:44 INFO TaskSchedulerImpl: Removed TaskSet 18.0, whose tasks have all completed, from pool
18/07/30 03:03:44 INFO DAGScheduler: ShuffleMapStage 18 (distinct at RDD_ASSIGNMENT.scala:88) finished in 0.026 s
Compilation completed successfully in 4 s 189 ms (7 minutes ago)
```

# RDD DEEP DIVE

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:44 INFO BlockManagerInfo: Added broadcast_20_piece0 in memory on LAPTOP-UCU4FHT9:65486 (size: 2.6 KB, free: 899.7 MB)
18/07/30 03:03:44 INFO SparkContext: Created broadcast 20 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:44 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 20 (MapPartitionsRDD[27] at mapValues at RDD_ASSIGNMENT.scala:88) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:44 INFO TaskSetManager: Starting task 0.0 in stage 20.0 (TID 19, localhost, executor driver, partition 0, ANY, 7649 bytes)
18/07/30 03:03:44 INFO Executor: Running task 0.0 in stage 20.0 (TID 19)
18/07/30 03:03:44 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:44 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/30 03:03:44 INFO Executor: Finished task 0.0 in stage 20.0 (TID 19). 1052 bytes result sent to driver
18/07/30 03:03:44 INFO TaskSetManager: Finished task 0.0 in stage 20.0 (TID 19) in 12 ms on localhost (executor driver) (1/1)
18/07/30 03:03:44 INFO TaskSchedulerImpl: Removed TaskSet 20.0, whose tasks have all completed, from pool
18/07/30 03:03:44 INFO DAGScheduler: ResultStage 20 (foreach at RDD_ASSIGNMENT.scala:89) finished in 0.021 s
18/07/30 03:03:44 INFO DAGScheduler: Job 12 Finished: foreach at RDD_ASSIGNMENT.scala:89, took 0.099212 s
((grade-3,Mathew),45)
((grade-1,Andrew),131)
((grade-2,Mathew),142)
((grade-1,John),116)
((grade-1,Mark),168)
((grade-1,Silva),24)
((grade-2,Mark),35)
((grade-3,Silva),86)
((grade-2,John),74)
((grade-3,Andrew),70)
((grade-2,Andrew),77)
((grade-2,Silva),122)
18/07/30 03:03:44 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:93
18/07/30 03:03:44 INFO DAGScheduler: Registering RDD 28 (map at RDD_ASSIGNMENT.scala:90)
18/07/30 03:03:44 INFO DAGScheduler: Got job 13 (foreach at RDD_ASSIGNMENT.scala:93) with 1 output partitions
18/07/30 03:03:44 INFO DAGScheduler: Final stage: ResultStage 22 (foreach at RDD_ASSIGNMENT.scala:93)
18/07/30 03:03:44 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 21)
18/07/30 03:03:44 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 21)
18/07/30 03:03:44 INFO DAGScheduler: Submitting ShuffleMapStage 21 (MapPartitionsRDD[28] at map at RDD_ASSIGNMENT.scala:90), which has no missing parents
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_21 stored as values in memory (estimated size 5.3 KB, free 899.4 MB)
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_21_piece0 stored as bytes in memory (estimated size 2.3 KB, free 899.4 MB)
18/07/30 03:03:44 INFO BlockManagerInfo: Added broadcast_21_piece0 in memory on LAPTOP-UCU4FHT9:65486 (size: 2.3 KB, free: 899.7 MB)
18/07/30 03:03:44 INFO SparkContext: Created broadcast 21 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:44 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 21 (MapPartitionsRDD[28] at map at RDD_ASSIGNMENT.scala:90) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:44 INFO TaskSchedulerImpl: Adding task set 21.0 with 1 tasks
18/07/30 03:03:44 INFO TaskSetManager: Starting task 0.0 in stage 21.0 (TID 20, localhost, executor driver, partition 0, PROCESS_LOCAL, 7881 bytes)
18/07/30 03:03:44 INFO Executor: Running task 0.0 in stage 21.0 (TID 20)
Compilation completed successfully in 4 s 189 ms (7 minutes ago)
20:1 CRLF : UTF-8 : 03:10 AM 30-07-2018
```

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:44 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 22 (MapPartitionsRDD[30] at mapValues at RDD_ASSIGNMENT.scala:92) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:44 INFO TaskSchedulerImpl: Adding task set 22.0 with 1 tasks
18/07/30 03:03:44 INFO TaskSetManager: Starting task 0.0 in stage 22.0 (TID 21, localhost, executor driver, partition 0, ANY, 7649 bytes)
18/07/30 03:03:44 INFO Executor: Running task 0.0 in stage 22.0 (TID 21)
18/07/30 03:03:44 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:44 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/30 03:03:44 INFO Executor: Finished task 0.0 in stage 22.0 (TID 21). 1009 bytes result sent to driver
((grade-1,Andrew),3)
18/07/30 03:03:44 INFO TaskSetManager: Finished task 0.0 in stage 22.0 (TID 21) in 6 ms on localhost (executor driver) (1/1)
((grade-2,Mathew),3)
18/07/30 03:03:44 INFO TaskSchedulerImpl: Removed TaskSet 22.0, whose tasks have all completed, from pool
((grade-1,John),2)
((grade-1,Mark),2)
((grade-1,Silva),1)
((grade-2,Mark),2)
((grade-3,Silva),1)
((grade-2,John),1)
((grade-3,Andrew),2)
((grade-3,Mathew),1)
((grade-2,Silva),2)
((grade-2,Andrew),1)
18/07/30 03:03:44 INFO DAGScheduler: ResultStage 22 (foreach at RDD_ASSIGNMENT.scala:93) finished in 0.011 s
18/07/30 03:03:44 INFO DAGScheduler: Job 13 Finished: foreach at RDD_ASSIGNMENT.scala:93, took 0.038438 s
18/07/30 03:03:44 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:96
18/07/30 03:03:44 INFO DAGScheduler: Got job 14 (foreach at RDD_ASSIGNMENT.scala:96) with 1 output partitions
18/07/30 03:03:44 INFO DAGScheduler: Final stage: ResultStage 26 (foreach at RDD_ASSIGNMENT.scala:96)
18/07/30 03:03:44 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 25, ShuffleMapStage 23)
18/07/30 03:03:44 INFO DAGScheduler: Missing parents: List()
18/07/30 03:03:44 INFO DAGScheduler: Submitting ResultStage 26 (MapPartitionsRDD[33] at join at RDD_ASSIGNMENT.scala:95), which has no missing parents
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_23 stored as values in memory (estimated size 8.4 KB, free 899.4 MB)
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_23_piece0 stored as bytes in memory (estimated size 4.1 KB, free 899.4 MB)
18/07/30 03:03:44 INFO BlockManagerInfo: Added broadcast_23_piece0 in memory on LAPTOP-UCU4FHT9:65486 (size: 4.1 KB, free: 899.7 MB)
18/07/30 03:03:44 INFO SparkContext: Created broadcast 23 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:44 INFO TaskSchedulerImpl: Adding task set 26.0 with 1 tasks
18/07/30 03:03:44 INFO TaskSetManager: Starting task 0.0 in stage 26.0 (TID 22, localhost, executor driver, partition 0, ANY, 7880 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 26.0 (TID 22)
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
Compilation completed successfully in 4 s 189 ms (7 minutes ago)
20:1 CRLF : UTF-8 : 03:10 AM 30-07-2018
```



# RDD DEEP DIVE

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:44 INFO MemoryStore: Block broadcast_23_piece0 stored as bytes in memory (estimated size 4.1 KB, free 899.4 MB)
18/07/30 03:03:44 INFO BlockManagerInfo: Added broadcast_23_piece0 in memory on LAPTOP-UCD4PHN79:65486 (size: 4.1 KB, free: 899.7 MB)
18/07/30 03:03:44 INFO SparkContext: Created broadcast 23 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:44 INFO DAGSchedulerImpl: Submitting 1 missing tasks from ResultStage 26 (MapPartitionsRDD[33] at join at RDD_ASSIGNMENT.scala:95) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:44 INFO DAGSchedulerImpl: Adding task set 26.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 26.0 (TID 22, localhost, executor driver, partition 0, ANY, 7880 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 26.0 (TID 22)
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
((grade-1,Andrew),(131,3))
((grade-2,Mathew),(142,3))
((grade-1,John),(116,3))
((grade-1,Mark),(160,2))
((grade-1,Lisa),(24,1))
((grade-2,Mark),(35,2))
((grade-3,Lisa),(66,1))
((grade-3,John),(74,1))
((grade-3,Andrew),(70,2))
((grade-3,Mathew),(45,1))
((grade-2,Lisa),(122,2))
((grade-2,Andrew),(77,1))
18/07/30 03:03:45 INFO Executor: Finished task 0.0 in stage 26.0 (TID 22). 1095 bytes result sent to driver
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 26.0 (TID 22) in 32 ms on localhost (executor driver) (1/1)
18/07/30 03:03:45 INFO DAGSchedulerImpl: Removed TaskSet 26.0, whose tasks have all completed, from pool
18/07/30 03:03:45 INFO DAGScheduler: ResultStage 26 (foreach at RDD_ASSIGNMENT.scala:96) finished in 0.046 s
18/07/30 03:03:45 INFO DAGScheduler: Job 14 finished: foreach at RDD_ASSIGNMENT.scala:96, took 0.051452 s
18/07/30 03:03:45 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:99
18/07/30 03:03:45 INFO DAGScheduler: Got job 15 (foreach at RDD_ASSIGNMENT.scala:99) with 1 output partitions
18/07/30 03:03:45 INFO DAGScheduler: Final stage: ResultStage 30 (foreach at RDD_ASSIGNMENT.scala:99)
18/07/30 03:03:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27, ShuffleMapStage 29)
18/07/30 03:03:45 INFO DAGScheduler: Missing parents: List()
18/07/30 03:03:45 INFO DAGScheduler: Submitting ResultStage 30 (MapPartitionsRDD[34] at map at RDD_ASSIGNMENT.scala:98), which has no missing parents
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_24_piece0 stored as values in memory (estimated size 8.5 KB, free 899.4 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_24_piece0 in memory on LAPTOP-UCD4PHN79:65486 (size: 4.1 KB, free: 899.6 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 24 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 30 (MapPartitionsRDD[34] at map at RDD_ASSIGNMENT.scala:98) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO DAGSchedulerImpl: Adding task set 30.0 with 1 tasks
Compilation completed successfully in 4 s 189 ms (7 minutes ago)
```

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_24_piece0 in memory on LAPTOP-UCD4PHN79:65486 (size: 4.1 KB, free: 899.6 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 24 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 30 (MapPartitionsRDD[34] at map at RDD_ASSIGNMENT.scala:98) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO DAGSchedulerImpl: Adding task set 30.0 with 1 tasks
18/07/30 03:03:45 INFO DAGSchedulerImpl: Starting task 0.0 in stage 30.0 (TID 23, localhost, executor driver, partition 0, ANY, 7880 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 30.0 (TID 23)
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
Average Of Student: ((grade-1,Andrew),43)
Average Of Student: ((grade-2,Mathew),47)
Average Of Student: ((grade-1,John),38)
Average Of Student: ((grade-1,Mark),84)
Average Of Student: ((grade-1,Lisa),24)
Average Of Student: ((grade-2,Mark),17)
Average Of Student: ((grade-3,Lisa),86)
Average Of Student: ((grade-2,John),74)
Average Of Student: ((grade-3,Andrew),35)
Average Of Student: ((grade-3,Mathew),45)
Average Of Student: ((grade-2,Lisa),61)
Average Of Student: ((grade-2,Andrew),77)
18/07/30 03:03:45 INFO Executor: Finished task 0.0 in stage 30.0 (TID 23). 1095 bytes result sent to driver
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 30.0 (TID 23) in 30 ms on localhost (executor driver) (1/1)
18/07/30 03:03:45 INFO DAGSchedulerImpl: Removed TaskSet 30.0, whose tasks have all completed, from pool
18/07/30 03:03:45 INFO DAGScheduler: ResultStage 30 (foreach at RDD_ASSIGNMENT.scala:98) finished in 0.043 s
18/07/30 03:03:45 INFO DAGScheduler: Job 15 finished: foreach at RDD_ASSIGNMENT.scala:99, took 0.048416 s
18/07/30 03:03:45 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:104
18/07/30 03:03:45 INFO DAGScheduler: Registering RDD 35 (map at RDD_ASSIGNMENT.scala:102)
18/07/30 03:03:45 INFO DAGScheduler: Got job 16 (foreach at RDD_ASSIGNMENT.scala:104) with 1 output partitions
18/07/30 03:03:45 INFO DAGScheduler: Final stage: ResultStage 32 (foreach at RDD_ASSIGNMENT.scala:104)
18/07/30 03:03:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 31)
18/07/30 03:03:45 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 31)
18/07/30 03:03:45 INFO DAGScheduler: Submitting ShuffleMapStage 31 (MapPartitionsRDD[35] at map at RDD_ASSIGNMENT.scala:102), which has no missing parents
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_25_piece0 stored as values in memory (estimated size 5.6 KB, free 899.4 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_25_piece0 in memory on LAPTOP-UCD4PHN79:65486 (size: 3.1 KB, free: 899.6 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 25 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 31 (MapPartitionsRDD[35] at map at RDD_ASSIGNMENT.scala:102) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO DAGSchedulerImpl: Adding task set 31.0 with 1 tasks
Compilation completed successfully in 4 s 189 ms (7 minutes ago)
```

# RDD DEEP DIVE

```
RDDEspark [C:\Users\Shruthi\IdeaProjects\RDDDEspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDDEspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDDEspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:45 INFO DAGScheduler: running: Set()
18/07/30 03:03:45 INFO DAGScheduler: waiting: Set(ResultStage 32)
18/07/30 03:03:45 INFO DAGScheduler: failed: Set()
18/07/30 03:03:45 INFO DAGScheduler: Submitting ResultStage 32 (MapPartitionsRDD[37] at mapValues at RDD_ASSIGNMENT.scala:103), which has no missing parents
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_36 stored as values in memory (estimated size 6.7 KB, free 899.4 MB)
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_36_piece0 stored as bytes in memory (estimated size 3.5 KB, free 899.4 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_36_piece0 in memory on LAPTOP-UCU4PH79:65486 (size: 3.5 KB, free: 899.4 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 26 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 32 (MapPartitionsRDD[37] at mapValues at RDD_ASSIGNMENT.scala:103) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSchedulerImpl: Adding task set 32.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 32.0 (TID 25, localhost, executor driver, partition 0, ANY, 7649 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 32.0 (TID 25)
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/30 03:03:45 INFO Executor: Finished task 0.0 in stage 32.0 (TID 25). 1095 bytes result sent to driver
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 32.0 (TID 25) in 15 ms on localhost (executor driver) (1/1)
18/07/30 03:03:45 INFO TaskSchedulerImpl: Removed TaskSet 32.0, whose tasks have all completed, from pool
18/07/30 03:03:45 INFO DAGScheduler: ResultStage 32 (foreach at RDD_ASSIGNMENT.scala:104) finished in 0.020 s
18/07/30 03:03:45 INFO DAGScheduler: Job 16 finished: foreach at RDD_ASSIGNMENT.scala:104, took 0.054520 s
18/07/30 03:03:45 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:104
18/07/30 03:03:45 INFO DAGScheduler: Registering RDD 36 (map at RDD_ASSIGNMENT.scala:105)
18/07/30 03:03:45 INFO DAGScheduler: Got job 17 (foreach at RDD_ASSIGNMENT.scala:106) with 1 output partitions
18/07/30 03:03:45 INFO DAGScheduler: Final stage: ResultStage 34 (foreach at RDD_ASSIGNMENT.scala:106)
18/07/30 03:03:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 33)
18/07/30 03:03:45 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 33)
18/07/30 03:03:45 INFO DAGScheduler: Submitting ShuffleMapStage 33 (MapPartitionsRDD[38] at map at RDD_ASSIGNMENT.scala:105), which has no missing parents
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_27 stored as values in memory (estimated size 5.6 KB, free 899.4 MB)
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_27_piece0 stored as bytes in memory (estimated size 3.1 KB, free 899.4 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_27_piece0 in memory on LAPTOP-UCU4PH79:65486 (size: 3.1 KB, free: 899.4 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 27 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 33 (MapPartitionsRDD[38] at map at RDD_ASSIGNMENT.scala:105) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSchedulerImpl: Adding task set 33.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 33.0 (TID 26, localhost, executor driver, partition 0, PROCESS_LOCAL, 7881 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 33.0 (TID 26)
18/07/30 03:03:45 INFO HadoopRDD: Input split: file:/C:/Users/Shruthi/Desktop/19_Dataset.txt:0+624
18/07/30 03:03:45 INFO Executor: Finished task 0.0 in stage 33.0 (TID 26). 898 bytes result sent to driver
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 33.0 (TID 26) in 47 ms on localhost (executor driver) (1/1)
Compilation completed successfully in 4 s 189 ms (7 minutes ago)
```

```
RDDEspark [C:\Users\Shruthi\IdeaProjects\RDDDEspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDDEspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDDEspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_28_piece0 stored as bytes in memory (estimated size 3.5 KB, free 899.3 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_28_piece0 in memory on LAPTOP-UCU4PH79:65486 (size: 3.5 KB, free: 899.3 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 28 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 34 (MapPartitionsRDD[40] at mapValues at RDD_ASSIGNMENT.scala:105) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSchedulerImpl: Adding task set 34.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 34.0 (TID 27, localhost, executor driver, partition 0, ANY, 7649 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 34.0 (TID 27)
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/30 03:03:45 INFO Executor: Finished task 0.0 in stage 34.0 (TID 27). 1138 bytes result sent to driver
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 34.0 (TID 27) in 25 ms on localhost (executor driver) (1/1)
18/07/30 03:03:45 INFO TaskSchedulerImpl: Removed TaskSet 34.0, whose tasks have all completed, from pool
18/07/30 03:03:45 INFO DAGScheduler: ResultStage 34 (foreach at RDD_ASSIGNMENT.scala:106) finished in 0.035 s
18/07/30 03:03:45 INFO DAGScheduler: Job 17 finished: foreach at RDD_ASSIGNMENT.scala:106, took 0.056936 s
18/07/30 03:03:45 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:106
18/07/30 03:03:45 INFO DAGScheduler: Got job 18 (foreach at RDD_ASSIGNMENT.scala:106) with 1 output partitions
18/07/30 03:03:45 INFO DAGScheduler: Final stage: ResultStage 37 (foreach at RDD_ASSIGNMENT.scala:106)
18/07/30 03:03:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 35, ShuffleMapStage 36)
18/07/30 03:03:45 INFO DAGScheduler: Missing parents: List()
18/07/30 03:03:45 INFO DAGScheduler: Submitting ResultStage 37 (MapPartitionsRDD[44] at map at RDD_ASSIGNMENT.scala:108), which has no missing parents
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_29 stored as values in memory (estimated size 8.2 KB, free 899.3 MB)
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 526
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 577
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 459
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 618
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 379
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 621
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 555
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 400
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 567
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 578
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 507
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 631
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 350
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 623
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 560
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 483
Compilation completed successfully in 4 s 189 ms (8 minutes ago)
```



# RDD DEEP DIVE

```
Run: RDD_ASSIGNMENT
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 552
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 529
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 478
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 469
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 425
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 455
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 486
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 474
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 689
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 682
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 351
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 645
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 675
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 430
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 395
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 378
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 357
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 625
18/07/30 03:03:45 INFO TaskScheduler: Finished task 0.0 in stage 37.0 (STD 28) in 48 ms on localhost (executor driver) (1/1)
18/07/30 03:03:45 INFO TaskSchedulerImpl: Removed TaskSet 37.0, whose tasks have all completed, from pool
Average for Subject:maths==>46
Average for Subject:science==>38
18/07/30 03:03:45 INFO DAGScheduler: ResultStage 37 (foreach at RDD_ASSIGNMENT.scala:105) finished in 0.055 s
18/07/30 03:03:45 INFO BlockManagerInfo: Removed broadcast_25_piece0 on LAPTOP-UCU4PMT9:65486 in memory (size: 3.1 KB, free: 899.6 MB)
18/07/30 03:03:45 INFO DAGScheduler: Job 18 finished: foreach at RDD_ASSIGNMENT.scala:105, took 0.064438 s
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 417
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 444
18/07/30 03:03:45 INFO BlockManagerInfo: Removed broadcast_20_piece0 on LAPTOP-UCU4PMT9:65486 in memory (size: 2.6 KB, free: 899.6 MB)
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 546
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 561
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 481
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 538
18/07/30 03:03:45 INFO BlockManagerInfo: Removed broadcast_19_piece0 on LAPTOP-UCU4PMT9:65486 in memory (size: 2.3 KB, free: 899.7 MB)
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 506
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 580
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 575
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 616
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 452
18/07/30 03:03:45 INFO ContextCleaner: Cleaned accumulator 452

Compilation completed successfully in 4 s 189 ms (8 minutes ago)
```

# RDD DEEP DIVE

## Problem Statement 3:

Are there any students in the college that satisfy the below criteria:

1. Average score per student\_name across all grades is same as average score per student\_name per grade

```
117 val noOfStudent1 = tuple2RDD.map(x=>{(x(1),x(2)),1}).groupByKey().mapValues(x=>x.sum)
118 noOfStudent1.foreach(x=>println(x))
119 //Join these RDD and calculate the average
120 val avg3 = sum4.join(noOfStudent1).map(x=>{(x._1,(x._2._1/x._2._2))})
121 avg3.foreach(x=>println("Average for : "+x._1+"====>"+x._2))
122
123 //5. For all students in grade=2, how many have average score greater than 50?
124 val grade2Students = averageOfEachStd.filter(x=>x._1._1.equalsIgnoreCase("grade=2") && x._2>50)
125 grade2Students.foreach(x=>println(x))
126
127
128 //Problem Statement 3:
129 //1. Average score per student_name across all grades is same as average score per student_name per grade
130 //Hint - Use Intersection Property
131 val sumPerStdName = tuple2RDD.map(x=>{(x(1),x(3)).toInt}).groupByKey().mapValues(y=>y.sum)
132 sumPerStdName.foreach(x=>println(x))
133 val countOfStdName = tuple2RDD.map(x=>{(x(0),1)}).groupByKey().mapValues(y=>y.sum)
134 countOfStdName.foreach(x=>println(x))
135 //Join these two RDDs to get average per student_name
136 val avgStdName = sumPerStdName.join(countOfStdName).map(x=>{(x._1,x._2._1/x._2._2)})
137 //The RDD containing average for student_name per grade is averageOfEachStd now intersect this RDD with avgStdName
138 val avgPerGrade = averageOfEachStd.map(x=>{(x._1,x._2,x._2)})
139 avgStdName.foreach(x=>println("Average per student_name="+x))
140 avgPerGrade.foreach(x=>println("Average per student_name & Grade="+x))
141 val commonTds = avgStdName.intersection(avgPerGrade)
142 commonTds.foreach(x=>println("Students having same average per name and per grade="+x))
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
```

```
Run: RDD_ASSIGNMENT x
Process finished with exit code 0

Compilation completed successfully in 4 s 189 ms (5 minutes ago)
```

```
18/07/30 03:03:45 INFO DAGScheduler: waiting: Set(ResultStage 39)
18/07/30 03:03:45 INFO DAGScheduler: failed: Set()
18/07/30 03:03:45 INFO DAGScheduler: Submitting ResultStage 39 (MapPartitionsRDD[47] at mapValues at RDD_ASSIGNMENT.scala:114), which has no missing parents
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_31 stored as values in memory (estimated size 6.4 KB, free 899.4 MB)
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_31_piece0 stored as bytes in memory (estimated size 3.3 KB, free 389.4 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_31_piece0 in memory on LAPTOP-UNCMPT9:65486 (size: 3.3 KB, free: 899.7 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 31 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 39 (MapPartitionsRDD[47] at mapValues at RDD_ASSIGNMENT.scala:114) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSetManagerImpl: Adding task set 39.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 39.0 (TID 30, localhost, executor driver, partition 0, ANY, 7649 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 39.0 (TID 30)
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 39.0 (TID 30), 1052 bytes result sent to driver
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 39.0 (TID 30) in 8 ms on localhost (executor driver) (1/1)
18/07/30 03:03:45 INFO TaskSetManagerImpl: Removed TaskSet 39.0, whose tasks have all completed, from pool
18/07/30 03:03:45 INFO DAGScheduler: ResultStage 39 (foreach at RDD_ASSIGNMENT.scala:115) finished in 0.017 s
18/07/30 03:03:45 INFO SparkContext: Job 19 finished: foreach at RDD_ASSIGNMENT.scala:115, took 0.002869 s
18/07/30 03:03:45 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:118
18/07/30 03:03:45 INFO DAGScheduler: Registering RDD 40 (map at RDD_ASSIGNMENT.scala:117)
18/07/30 03:03:45 INFO DAGScheduler: Got job 20 (foreach at RDD_ASSIGNMENT.scala:118) with 1 output partitions
18/07/30 03:03:45 INFO DAGScheduler: Final stage: ResultStage 41 (foreach at RDD_ASSIGNMENT.scala:118)
18/07/30 03:03:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 40)
18/07/30 03:03:45 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 40)
18/07/30 03:03:45 INFO DAGScheduler: Submitting ShuffleMapStage 40 (MapPartitionsRDD[48] at map at RDD_ASSIGNMENT.scala:117), which has no missing parents
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_32 stored as values in memory (estimated size 5.3 KB, free 899.4 MB)
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_32_piece0 stored as bytes in memory (estimated size 2.9 KB, free 389.4 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_32_piece0 in memory on LAPTOP-UNCMPT9:65486 (size: 2.9 KB, free: 899.7 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 32 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 40 (MapPartitionsRDD[48] at map at RDD_ASSIGNMENT.scala:117) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSetManagerImpl: Adding task set 40.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 40.0 (TID 31, localhost, executor driver, partition 0, PROCESS_LOCAL, 7681 bytes)
```

```
Compilation completed successfully in 4 s 189 ms (8 minutes ago)
```

# RDD DEEP DIVE

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:45 INFO DAGScheduler: failed: Set()
18/07/30 03:03:45 INFO DAGScheduler: Submitting ResultStage 41 (MapPartitionsRDD[50] at mapValues at RDD_ASSIGNMENT.scala:117), which has no missing parents
18/07/30 03:03:45 INFO TaskSetManagerImpl: Removed TaskSet 40.0, whose tasks have all completed, from pool
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_33 stored as values in memory (estimated size 6.4 KB, free 899.4 MB)
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_33_piece0 stored as bytes in memory (estimated size 3.3 KB, free 899.4 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_33_piece0 in memory on LAPTOP-UCM4PHN9:65486 (size: 3.3 KB, free: 899.7 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 33 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 41 (MapPartitionsRDD[50] at mapValues at RDD_ASSIGNMENT.scala:117) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSchedulerImpl: Adding task set 41.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 41.0 (TID 32, localhost, executor driver, partition 0, ANY, 7649 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 41.0 (TID 32)
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/30 03:03:45 INFO Executor: Finished task 0.0 in stage 41.0 (TID 32). 1052 bytes result sent to driver
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 41.0 (TID 32) in 6 ms on localhost (executor driver) (1/1)
18/07/30 03:03:45 INFO TaskSchedulerImpl: Removed TaskSet 41.0, whose tasks have all completed, from pool
18/07/30 03:03:45 INFO DAGScheduler: ResultStage 41 (foreach at RDD_ASSIGNMENT.scala:118) finished in 0.013 s
18/07/30 03:03:45 INFO DAGScheduler: Job 20 finished: foreach at RDD_ASSIGNMENT.scala:118, took 0.072736 s
(history,grade-2),1)
(maths,grade-1),4)
(science,grade-3),3)
(science,grade-1),2)
(science,grade-2),3)
(history,grade-1),3)
(maths,grade-2),3)
18/07/30 03:03:45 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:121
18/07/30 03:03:45 INFO DAGScheduler: Got job 21 (foreach at RDD_ASSIGNMENT.scala:121) with 1 output partitions
18/07/30 03:03:45 INFO DAGScheduler: Final stage: ResultStage 44 (foreach at RDD_ASSIGNMENT.scala:121)
18/07/30 03:03:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 42, ShuffleMapStage 43)
18/07/30 03:03:45 INFO DAGScheduler: Missing parents: List()
18/07/30 03:03:45 INFO DAGScheduler: Submitting ResultStage 44 (MapPartitionsRDD[54] at map at RDD_ASSIGNMENT.scala:120), which has no missing parents
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_34 stored as values in memory (estimated size 7.0 KB, free 899.4 MB)
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_34_piece0 stored as bytes in memory (estimated size 3.0 KB, free 899.4 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_34_piece0 in memory on LAPTOP-UCM4PHN9:65486 (size: 3.0 KB, free: 899.7 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 34 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 44 (MapPartitionsRDD[54] at map at RDD_ASSIGNMENT.scala:120) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSchedulerImpl: Adding task set 44.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 44.0 (TID 33, localhost, executor driver, partition 0, ANY, 7886 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 44.0 (TID 33)
Compilation completed successfully in 4 s 189 ms (8 minutes ago)
20:1 CRLF UTF-8 03:11 AM 30-07-2018
```

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:45 INFO DAGScheduler: Missing parents: List()
18/07/30 03:03:45 INFO DAGScheduler: Submitting ResultStage 44 (MapPartitionsRDD[54] at map at RDD_ASSIGNMENT.scala:120), which has no missing parents
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_34 stored as values in memory (estimated size 7.0 KB, free 899.4 MB)
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_34_piece0 stored as bytes in memory (estimated size 3.0 KB, free 899.4 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_34_piece0 in memory on LAPTOP-UCM4PHN9:65486 (size: 3.0 KB, free: 899.7 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 34 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 44 (MapPartitionsRDD[54] at map at RDD_ASSIGNMENT.scala:120) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSchedulerImpl: Adding task set 44.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 44.0 (TID 33, localhost, executor driver, partition 0, ANY, 7886 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 44.0 (TID 33)
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
Average for : (history,grade-2) ==>79
Average for : (history,grade-3) ==>86
Average for : (maths,grade-1) ==>46
Average for : (science,grade-3) ==>38
Average for : (science,grade-1) ==>30
Average for : (science,grade-2) ==>30
Average for : (history,grade-1) ==>51
Average for : (maths,grade-2) ==>48
18/07/30 03:03:45 INFO Executor: Finished task 0.0 in stage 44.0 (TID 33). 1095 bytes result sent to driver
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 44.0 (TID 33) in 25 ms on localhost (executor driver) (1/1)
18/07/30 03:03:45 INFO TaskSchedulerImpl: Removed TaskSet 44.0, whose tasks have all completed, from pool
18/07/30 03:03:45 INFO DAGScheduler: ResultStage 44 (foreach at RDD_ASSIGNMENT.scala:121) finished in 0.036 s
18/07/30 03:03:45 INFO DAGScheduler: Job 21 finished: foreach at RDD_ASSIGNMENT.scala:121, took 0.039901 s
18/07/30 03:03:45 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:125
18/07/30 03:03:45 INFO DAGScheduler: Got job 22 (foreach at RDD_ASSIGNMENT.scala:125) with 1 output partitions
18/07/30 03:03:45 INFO DAGScheduler: Final stage: ResultStage 48 (foreach at RDD_ASSIGNMENT.scala:125)
18/07/30 03:03:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 45, ShuffleMapStage 47)
18/07/30 03:03:45 INFO DAGScheduler: Missing parents: List()
18/07/30 03:03:45 INFO DAGScheduler: Submitting ResultStage 48 (MapPartitionsRDD[55] at filter at RDD_ASSIGNMENT.scala:124), which has no missing parents
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_35 stored as values in memory (estimated size 8.7 KB, free 899.4 MB)
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_35_piece0 stored as bytes in memory (estimated size 4.2 KB, free 899.4 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_35_piece0 in memory on LAPTOP-UCM4PHN9:65486 (size: 4.2 KB, free: 899.7 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 35 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 48 (MapPartitionsRDD[55] at filter at RDD_ASSIGNMENT.scala:124) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSchedulerImpl: Adding task set 48.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 48.0 (TID 34, localhost, executor driver, partition 0, ANY, 7886 bytes)
Compilation completed successfully in 4 s 189 ms (8 minutes ago)
20:1 CRLF UTF-8 03:11 AM 30-07-2018
```



# RDD DEEP DIVE

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_35 stored as values in memory (estimated size 0.7 KB, free 899.4 MB)
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_35_piece0 stored as bytes in memory (estimated size 4.2 KB, free 899.4 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_35_piece0 in memory on LAPTOP-UCU4PH99:65486 (size: 4.2 KB, free: 899.7 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 35 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 48 (MapPartitionsRDD[55] at filter at RDD_ASSIGNMENT.scala:124) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSchedulerImpl: Adding task set 48.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 48.0 (TID 34, localhost, executor driver, partition 0, ANY, 7888 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 48.0 (TID 34)
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO Executor: Finished task 0.0 in stage 48.0 (TID 34). 1095 bytes result sent to driver
((grade-2,John),74)
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 48.0 (TID 34) in 18 ms on localhost (executor driver) (1/1)
((grade-2,Lisa),61)
((grade-2,Andrew),77)
18/07/30 03:03:45 INFO TaskSchedulerImpl: Removed TaskSet 48.0, whose tasks have all completed, from pool
18/07/30 03:03:45 INFO DAGScheduler: ResultStage 48 (foreach at RDD_ASSIGNMENT.scala:125) Finished in 0.026 s
18/07/30 03:03:45 INFO DAGScheduler: Job 22 Finished: foreach at RDD_ASSIGNMENT.scala:125, took 0.03103 s
18/07/30 03:03:45 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:132
18/07/30 03:03:45 INFO DAGScheduler: Registering RDD 56 (map at RDD_ASSIGNMENT.scala:131)
18/07/30 03:03:45 INFO DAGScheduler: Got job 23 (foreach at RDD_ASSIGNMENT.scala:132) with 1 output partitions
18/07/30 03:03:45 INFO DAGScheduler: Final stage: ResultStage 50 (foreach at RDD_ASSIGNMENT.scala:132)
18/07/30 03:03:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 49)
18/07/30 03:03:45 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 49)
18/07/30 03:03:45 INFO DAGScheduler: Submitting ShuffleMapStage 49 (MapPartitionsRDD[56] at map at RDD_ASSIGNMENT.scala:131), which has no missing parents
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_36 stored as values in memory (estimated size 5.6 KB, free 899.4 MB)
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_36_piece0 stored as bytes in memory (estimated size 3.1 KB, free 899.4 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_36_piece0 in memory on LAPTOP-UCU4PH99:65486 (size: 3.1 KB, free: 899.6 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 36 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 49 (MapPartitionsRDD[56] at map at RDD_ASSIGNMENT.scala:131) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSchedulerImpl: Adding task set 49.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 49.0 (TID 35, localhost, executor driver, partition 0, PROCESS_LOCAL, 7881 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 49.0 (TID 35)
18/07/30 03:03:45 INFO HadoopRDD: Input split: file:/C:/Users/Shruthi/Desktop/19_Dataset.txt:0+624
18/07/30 03:03:45 INFO Executor: Finished task 0.0 in stage 49.0 (TID 35). 941 bytes result sent to driver
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 49.0 (TID 35) in 26 ms on localhost (executor driver) (1/1)
18/07/30 03:03:45 INFO TaskSchedulerImpl: Removed TaskSet 49.0, whose tasks have all completed, from pool
18/07/30 03:03:45 INFO DAGScheduler: ShuffleMapStage 49 (map at RDD_ASSIGNMENT.scala:131) finished in 0.031 s
Compilation completed successfully in 4 s 189 ms (8 minutes ago)
```

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_37 stored as values in memory (estimated size 6.7 KB, free 899.4 MB)
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_37_piece0 stored as bytes in memory (estimated size 3.5 KB, free 899.4 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_37_piece0 in memory on LAPTOP-UCU4PH99:65486 (size: 3.5 KB, free: 899.6 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 37 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 50 (MapPartitionsRDD[58] at mapValues at RDD_ASSIGNMENT.scala:131) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSchedulerImpl: Adding task set 50.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 50.0 (TID 36, localhost, executor driver, partition 0, ANY, 7649 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 50.0 (TID 36)
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
(Mark,203)
(Andrew,278)
(Mathew,242)
(John,190)
(Lisa,232)
18/07/30 03:03:45 INFO Executor: Finished task 0.0 in stage 50.0 (TID 36). 1095 bytes result sent to driver
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 50.0 (TID 36) in 12 ms on localhost (executor driver) (1/1)
18/07/30 03:03:45 INFO TaskSchedulerImpl: Removed TaskSet 50.0, whose tasks have all completed, from pool
18/07/30 03:03:45 INFO DAGScheduler: ResultStage 50 (foreach at RDD_ASSIGNMENT.scala:132) Finished in 0.028 s
18/07/30 03:03:45 INFO DAGScheduler: Job 23 Finished: foreach at RDD_ASSIGNMENT.scala:132, took 0.058467 s
18/07/30 03:03:45 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:134
18/07/30 03:03:45 INFO DAGScheduler: Registering RDD 59 (map at RDD_ASSIGNMENT.scala:133)
18/07/30 03:03:45 INFO DAGScheduler: Got job 24 (foreach at RDD_ASSIGNMENT.scala:134) with 1 output partitions
18/07/30 03:03:45 INFO DAGScheduler: Final stage: ResultStage 52 (foreach at RDD_ASSIGNMENT.scala:134)
18/07/30 03:03:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 51)
18/07/30 03:03:45 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 51)
18/07/30 03:03:45 INFO DAGScheduler: Submitting ShuffleMapStage 51 (MapPartitionsRDD[59] at map at RDD_ASSIGNMENT.scala:133), which has no missing parents
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_38 stored as values in memory (estimated size 5.6 KB, free 899.4 MB)
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_38_piece0 stored as bytes in memory (estimated size 3.1 KB, free 899.4 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_38_piece0 in memory on LAPTOP-UCU4PH99:65486 (size: 3.1 KB, free: 899.6 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 38 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 51 (MapPartitionsRDD[59] at map at RDD_ASSIGNMENT.scala:133) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSchedulerImpl: Adding task set 51.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 51.0 (TID 37, localhost, executor driver, partition 0, PROCESS_LOCAL, 7881 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 51.0 (TID 37)
18/07/30 03:03:45 INFO HadoopRDD: Input split: file:/C:/Users/Shruthi/Desktop/19_Dataset.txt:0+624
18/07/30 03:03:45 INFO Executor: Finished task 0.0 in stage 51.0 (TID 37). 941 bytes result sent to driver
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 51.0 (TID 37) in 33 ms on localhost (executor driver) (1/1)
18/07/30 03:03:45 INFO TaskSchedulerImpl: Removed TaskSet 51.0, whose tasks have all completed, from pool
18/07/30 03:03:45 INFO DAGScheduler: ShuffleMapStage 51 (map at RDD_ASSIGNMENT.scala:133) finished in 0.039 s
Compilation completed successfully in 4 s 189 ms (8 minutes ago)
```



# RDD DEEP DIVE

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:45 INFO SparkContext: Created broadcast 39 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 52 (MapPartitionsRDD[61] at mapValues at RDD_ASSIGNMENT.scala:133) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSchedulerImpl: Adding task set 52.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 52.0 (TID 38, localhost, executor driver, partition 0, ANY, 7649 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 52.0 (TID 38)
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/30 03:03:45 INFO Executor: Finished task 0.0 in stage 52.0 (TID 38). 1095 bytes result sent to driver
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 52.0 (TID 38) in 22 ms on localhost (executor driver) (1/1)
18/07/30 03:03:45 INFO TaskSchedulerImpl: Removed TaskSet 52.0, whose tasks have all completed, from pool
18/07/30 03:03:45 INFO DAGScheduler: ResultStage 52 (foreach at RDD_ASSIGNMENT.scala:134) finished in 0.029 s
18/07/30 03:03:45 INFO DAGScheduler: Job 24 finished: foreach at RDD_ASSIGNMENT.scala:134, took 0.075570 s
(Mark,4)
(Andrew,6)
(Mathew,4)
(John,4)
(Lisa,4)
18/07/30 03:03:45 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:139
18/07/30 03:03:45 INFO DAGScheduler: Got job 25 (foreach at RDD_ASSIGNMENT.scala:139) with 1 output partitions
18/07/30 03:03:45 INFO DAGScheduler: Final stage: ResultStage 55 (foreach at RDD_ASSIGNMENT.scala:139)
18/07/30 03:03:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 53, ShuffleMapStage 54)
18/07/30 03:03:45 INFO DAGScheduler: Missing parents: List()
18/07/30 03:03:45 INFO DAGScheduler: Submitting ResultStage 55 (MapPartitionsRDD[65] at map at RDD_ASSIGNMENT.scala:136), which has no missing parents
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_40 stored as values in memory (estimated size 8.2 KB, free 899.3 MB)
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_40_piece0 stored as bytes in memory (estimated size 4.0 KB, free 899.3 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_40_piece0 in memory on LAPTOP-UC04PH99:5486 (size: 4.0 KB, free: 899.4 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 40 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 55 (MapPartitionsRDD[65] at map at RDD_ASSIGNMENT.scala:136) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSchedulerImpl: Adding task set 55.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 55.0 (TID 39, localhost, executor driver, partition 0, ANY, 7888 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 55.0 (TID 39)
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/30 03:03:45 INFO Executor: Finished task 0.0 in stage 55.0 (TID 39). 1138 bytes result sent to driver
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 55.0 (TID 39) in 25 ms on localhost (executor driver) (1/1)
18/07/30 03:03:45 INFO TaskSchedulerImpl: Removed TaskSet 55.0, whose tasks have all completed, from pool
18/07/30 03:03:45 INFO DAGScheduler: ResultStage 55 (foreach at RDD_ASSIGNMENT.scala:139) finished in 0.033 s
Average per student_Name (Mark,50)
Compilation completed successfully in 4 s 189 ms (8 minutes ago)
20:1 CRLF UTF-8 03:11 AM 30-07-2018
```

```
RDDspark [C:\Users\Shruthi\IdeaProjects\RDDspark] - ...src\main\scala\com\sparkrdd\assignment\RDD_ASSIGNMENT.scala [RDDspark] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
RDDspark src main scala com sparkrdd assignment RDD_ASSIGNMENT.scala
Run: RDD_ASSIGNMENT x
18/07/30 03:03:45 INFO DAGScheduler: Got job 25 (foreach at RDD_ASSIGNMENT.scala:139) with 1 output partitions
18/07/30 03:03:45 INFO DAGScheduler: Final stage: ResultStage 55 (foreach at RDD_ASSIGNMENT.scala:139)
18/07/30 03:03:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 53, ShuffleMapStage 54)
18/07/30 03:03:45 INFO DAGScheduler: Missing parents: List()
18/07/30 03:03:45 INFO DAGScheduler: Submitting ResultStage 55 (MapPartitionsRDD[65] at map at RDD_ASSIGNMENT.scala:136), which has no missing parents
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_40 stored as values in memory (estimated size 8.2 KB, free 899.3 MB)
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_40_piece0 stored as bytes in memory (estimated size 4.0 KB, free 899.3 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_40_piece0 in memory on LAPTOP-UC04PH99:5486 (size: 4.0 KB, free: 899.6 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 40 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 55 (MapPartitionsRDD[65] at map at RDD_ASSIGNMENT.scala:136) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSchedulerImpl: Adding task set 55.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 55.0 (TID 39, localhost, executor driver, partition 0, ANY, 7888 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 55.0 (TID 39)
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/30 03:03:45 INFO Executor: Finished task 0.0 in stage 55.0 (TID 39). 1138 bytes result sent to driver
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 55.0 (TID 39) in 25 ms on localhost (executor driver) (1/1)
18/07/30 03:03:45 INFO TaskSchedulerImpl: Removed TaskSet 55.0, whose tasks have all completed, from pool
18/07/30 03:03:45 INFO DAGScheduler: ResultStage 55 (foreach at RDD_ASSIGNMENT.scala:139) finished in 0.033 s
Average per student_Name (Mark,50)
Average per student_Name (Andrew,46)
Average per student_Name (Mathew,60)
Average per student_Name (John,47)
Average per student_Name (Lisa,58)
18/07/30 03:03:45 INFO DAGScheduler: Job 25 finished: foreach at RDD_ASSIGNMENT.scala:139, took 0.036026 s
18/07/30 03:03:45 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:140
18/07/30 03:03:45 INFO DAGScheduler: Got job 26 (foreach at RDD_ASSIGNMENT.scala:140) with 1 output partitions
18/07/30 03:03:45 INFO DAGScheduler: Final stage: ResultStage 59 (foreach at RDD_ASSIGNMENT.scala:140)
18/07/30 03:03:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 56, ShuffleMapStage 58)
18/07/30 03:03:45 INFO DAGScheduler: Missing parents: List()
18/07/30 03:03:45 INFO DAGScheduler: Submitting ResultStage 59 (MapPartitionsRDD[66] at map at RDD_ASSIGNMENT.scala:138), which has no missing parents
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_41 stored as values in memory (estimated size 8.6 KB, free 899.3 MB)
18/07/30 03:03:45 INFO MemoryStore: Block broadcast_41_piece0 stored as bytes in memory (estimated size 4.1 KB, free 899.3 MB)
18/07/30 03:03:45 INFO BlockManagerInfo: Added broadcast_41_piece0 in memory on LAPTOP-UC04PH99:5486 (size: 4.1 KB, free: 899.6 MB)
18/07/30 03:03:45 INFO SparkContext: Created broadcast 41 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 59 (MapPartitionsRDD[66] at map at RDD_ASSIGNMENT.scala:138) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSchedulerImpl: Adding task set 59.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSetManager: Starting task 0.0 in stage 59.0 (TID 40, localhost, executor driver, partition 0, ANY, 7888 bytes)
Compilation completed successfully in 4 s 189 ms (8 minutes ago)
20:1 CRLF UTF-8 03:12 AM 30-07-2018
```

# RDD DEEP DIVE

Run: RDD\_ASSIGNMENT

```
18/07/30 03:03:45 INFO SparkContext: Created broadcast 41 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 59 (MapPartitionsRDD[66] at map at RDD_ASSIGNMENT.scala:138) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:45 INFO TaskSchedulerImpl: Adding task set 59.0 with 1 tasks
18/07/30 03:03:45 INFO TaskSchedulerImpl: Starting task 0.0 in stage 59.0 (TID 40), localhost, executor driver, partition 0, ANV, 7888 bytes)
18/07/30 03:03:45 INFO Executor: Running task 0.0 in stage 59.0 (TID 40)
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/30 03:03:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
Average per student_Name & Grade (Andrew,43)
Average per student_Name & Grade (Mathew,47)
Average per student_Name & Grade (John,38)
Average per student_Name & Grade (Mark,84)
Average per student_Name & Grade (Lisa,24)
Average per student_Name & Grade (Mark,17)
Average per student_Name & Grade (Lisa,86)
Average per student_Name & Grade (John,74)
Average per student_Name & Grade (Andrew,35)
Average per student_Name & Grade (Mathew,45)
Average per student_Name & Grade (Lisa,61)
Average per student_Name & Grade (Andrew,77)
18/07/30 03:03:45 INFO Executor: Finished task 0.0 in stage 59.0 (TID 40). 1095 bytes result sent to driver
18/07/30 03:03:45 INFO TaskSetManager: Finished task 0.0 in stage 59.0 (TID 40) in 15 ms on localhost (executor driver) (1/1)
18/07/30 03:03:45 INFO TaskSchedulerImpl: Removed TaskSet 59.0, whose tasks have all completed, from pool
18/07/30 03:03:45 INFO DAGScheduler: ResultStage 59 (foreach at RDD_ASSIGNMENT.scala:140) finished in 0.022 s
18/07/30 03:03:46 INFO DAGScheduler: Job 26 finished: foreach at RDD_ASSIGNMENT.scala:140, took 0.024593 s
18/07/30 03:03:46 INFO SparkContext: Starting job: foreach at RDD_ASSIGNMENT.scala:140
18/07/30 03:03:46 INFO DAGScheduler: Registering RDD 68 (intersection at RDD_ASSIGNMENT.scala:141)
18/07/30 03:03:46 INFO DAGScheduler: Registering RDD 67 (intersection at RDD_ASSIGNMENT.scala:141)
18/07/30 03:03:46 INFO DAGScheduler: Got job 27 (foreach at RDD_ASSIGNMENT.scala:142) with 1 output partitions
18/07/30 03:03:46 INFO DAGScheduler: Final stage: ResultStage 67 (foreach at RDD_ASSIGNMENT.scala:142)
18/07/30 03:03:46 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 66, ShuffleMapStage 63)
18/07/30 03:03:46 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 66, ShuffleMapStage 63)
18/07/30 03:03:46 INFO DAGScheduler: Submitting ShuffleMapStage 63 (MapPartitionsRDD[68] at intersection at RDD_ASSIGNMENT.scala:141), which has no missing parents
18/07/30 03:03:46 INFO MemoryStore: Block broadcast_42_piece0 stored as values in memory (estimated size 8.7 KB, free 899.3 MB)
18/07/30 03:03:46 INFO MemoryStore: Block broadcast_42_piece0 stored as bytes in memory (estimated size 4.1 KB, free 899.3 MB)
18/07/30 03:03:46 INFO BlockManagerInfo: Added broadcast_42_piece0 in memory on LAPTOP-UCU4PHN9:5486 (size: 4.1 KB, free: 899.6 MB)
18/07/30 03:03:46 INFO SparkContext: Created broadcast 42 from broadcast at DAGScheduler.scala:1039
18/07/30 03:03:46 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 63 (MapPartitionsRDD[68] at intersection at RDD_ASSIGNMENT.scala:141) (first 15 tasks are for partitions Vector(0))
18/07/30 03:03:46 INFO TaskSchedulerImpl: Adding task set 68.0 with 1 tasks
```

Compilation completed successfully in 4 s 189 ms (8 minutes ago)

Spark RDD Example - Sj

laptop-ucu4fh9-4040/jobs/

Spark Jobs (7)

User: Shruthi  
Total Uptime: 10 s  
Scheduling Mode: FIFO  
Active Jobs: 1  
Completed Jobs: 23, only showing 19

Event Timeline

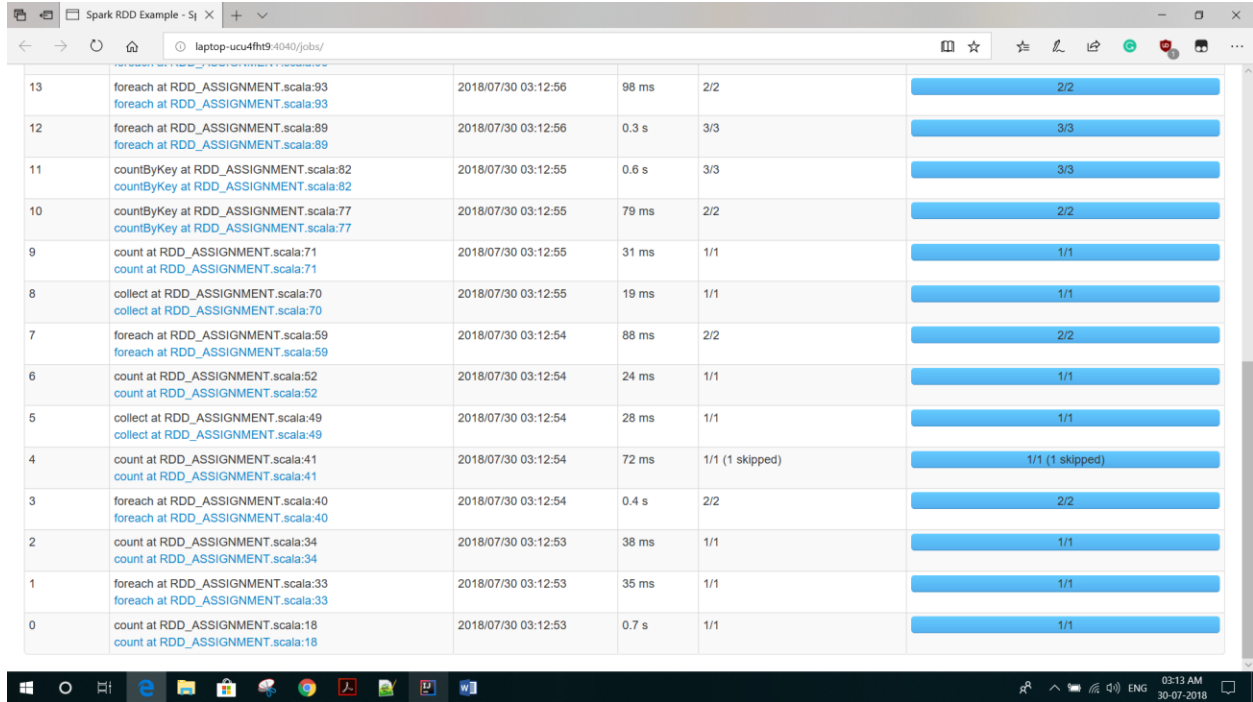
Active Jobs (1)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
19	foreach at RDD_ASSIGNMENT.scala:115 foreach at RDD_ASSIGNMENT.scala:115	2018/07/30 03:12:57 (kill)	0.6 s	0/2	0/2

Completed Jobs (23, only showing 19)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
18	foreach at RDD_ASSIGNMENT.scala:109 foreach at RDD_ASSIGNMENT.scala:109	2018/07/30 03:12:57	0.1 s	1/1 (2 skipped)	1/1 (2 skipped)
17	foreach at RDD_ASSIGNMENT.scala:106 foreach at RDD_ASSIGNMENT.scala:106	2018/07/30 03:12:57	0.2 s	2/2	2/2
16	foreach at RDD_ASSIGNMENT.scala:104 foreach at RDD_ASSIGNMENT.scala:104	2018/07/30 03:12:56	0.4 s	2/2	2/2
15	foreach at RDD_ASSIGNMENT.scala:99 foreach at RDD_ASSIGNMENT.scala:99	2018/07/30 03:12:56	28 ms	1/1 (3 skipped)	1/1 (3 skipped)
14	foreach at RDD_ASSIGNMENT.scala:96 foreach at RDD_ASSIGNMENT.scala:96	2018/07/30 03:12:56	49 ms	1/1 (3 skipped)	1/1 (3 skipped)
13	foreach at RDD_ASSIGNMENT.scala:93 foreach at RDD_ASSIGNMENT.scala:93	2018/07/30 03:12:56	98 ms	2/2	2/2

# RDD DEEP DIVE



13	foreach at RDD_ASSIGNMENT.scala:93 foreach at RDD_ASSIGNMENT.scala:93	2018/07/30 03:12:56	98 ms	2/2	2/2
12	foreach at RDD_ASSIGNMENT.scala:89 foreach at RDD_ASSIGNMENT.scala:89	2018/07/30 03:12:56	0.3 s	3/3	3/3
11	countByKey at RDD_ASSIGNMENT.scala:82 countByKey at RDD_ASSIGNMENT.scala:82	2018/07/30 03:12:55	0.6 s	3/3	3/3
10	countByKey at RDD_ASSIGNMENT.scala:77 countByKey at RDD_ASSIGNMENT.scala:77	2018/07/30 03:12:55	79 ms	2/2	2/2
9	count at RDD_ASSIGNMENT.scala:71 count at RDD_ASSIGNMENT.scala:71	2018/07/30 03:12:55	31 ms	1/1	1/1
8	collect at RDD_ASSIGNMENT.scala:70 collect at RDD_ASSIGNMENT.scala:70	2018/07/30 03:12:55	19 ms	1/1	1/1
7	foreach at RDD_ASSIGNMENT.scala:59 foreach at RDD_ASSIGNMENT.scala:59	2018/07/30 03:12:54	88 ms	2/2	2/2
6	count at RDD_ASSIGNMENT.scala:52 count at RDD_ASSIGNMENT.scala:52	2018/07/30 03:12:54	24 ms	1/1	1/1
5	collect at RDD_ASSIGNMENT.scala:49 collect at RDD_ASSIGNMENT.scala:49	2018/07/30 03:12:54	28 ms	1/1	1/1
4	count at RDD_ASSIGNMENT.scala:41 count at RDD_ASSIGNMENT.scala:41	2018/07/30 03:12:54	72 ms	1/1 (1 skipped)	1/1 (1 skipped)
3	foreach at RDD_ASSIGNMENT.scala:40 foreach at RDD_ASSIGNMENT.scala:40	2018/07/30 03:12:54	0.4 s	2/2	2/2
2	count at RDD_ASSIGNMENT.scala:34 count at RDD_ASSIGNMENT.scala:34	2018/07/30 03:12:53	38 ms	1/1	1/1
1	foreach at RDD_ASSIGNMENT.scala:33 foreach at RDD_ASSIGNMENT.scala:33	2018/07/30 03:12:53	35 ms	1/1	1/1
0	count at RDD_ASSIGNMENT.scala:18 count at RDD_ASSIGNMENT.scala:18	2018/07/30 03:12:53	0.7 s	1/1	1/1