01-Apache_Spark_Basics_Notes

February 22, 2025

0.1 Word Count Example

```
[1]: from pyspark.sql import SparkSession
```

```
[2]: spark = SparkSession.builder \
             .appName("Log Word Count") \
             .getOrCreate()
```

```
25/02/10 16:14:15 WARN SparkSession: Using an existing Spark session; only
runtime SQL configurations will take effect.
```

```
[3]: !hadoop fs -ls /log_data
```

```
Found 1 items
-rw-r--r--   2 lokesh hadoop   66622621 2025-02-10 16:10 /log_data/logfile.txt
```

```
[4]: hdfs_path = "/log_data/logfile.txt"
```

```
[5]: rdd = spark.sparkContext.textFile(hdfs_path)
```

```
[7]: rdd.take(5)
```

```
[7]: ['2025-02-01 17:54:46 - CRITICAL - Connection timeout on server server3',
      '2025-02-01 17:54:46 - INFO - Error in module moduleC: Null pointer exception',
      '2025-02-01 17:54:46 - ERROR - Disk space running low: 46% remaining',
      '2025-02-01 17:54:46 - INFO - Connection timeout on server server2',
      '2025-02-01 17:54:46 - ERROR - User 36 logged in']
```

```
[8]: rdd_split = rdd.map(lambda line: line.split("-"))
```

```
[10]: rdd_split.take(5)
```

```
[10]: [['2025',
        '02',
        '01 17:54:46 ',
        ' CRITICAL ',
        ' Connection timeout on server server3'],
       ['2025',
        '02',
        '01 17:54:46 ',
        ' INFO ',
        ' Error in module moduleC: Null pointer exception'],
       ['2025',
        '02',
        '01 17:54:46 ',
        ' ERROR ',
        ' Disk space running low: 46% remaining'],
       ['2025',
        '02',
        '01 17:54:46 ',
        ' INFO ',
        ' Connection timeout on server server2'],
       ['2025', '02', '01 17:54:46 ', ' ERROR ', ' User 36 logged in']]
```

```
[11]: flattened_rdd = rdd_split.flatMap(lambda x: [item.strip() for item in x if item.
       ↪strip() != ''])
```

```
[12]: flattened_rdd.take(5)
```

```
[12]: ['2025',
       '02',
       '01 17:54:46',
       'CRITICAL',
       'Connection timeout on server server3']
```

```
[13]: word_count = flattened_rdd.map(lambda word : (word, 1))
```

```
[14]: word_count_reduced = word_count.reduceByKey(lambda a, b : a + b)
```

```
[15]: word_count_reduced.collect()
```

```
[15]: [('2025', 1000000),
       ('INFO', 199567),
       ('User 36 logged in', 1971),
       ('User 100 logged in', 2019),
       ('DEBUG', 200535),
       ('User 48 logged in', 2031),
       ('Disk space running low: 6% remaining', 3974),
       ('User 58 logged in', 2060),
       ('User 34 logged in', 1966),
       ('User 22 logged in', 1982),
       ('User 8 logged in', 2002),
       ('Disk space running low: 50% remaining', 3963),
       ('Disk space running low: 32% remaining', 4045),
       ('Disk space running low: 29% remaining', 3961),
       ("File 'file1.txt' uploaded successfully", 66484),
       ('Disk space running low: 20% remaining', 3954),
       ('User 84 logged in', 2027),
       ('User 42 logged in', 1974),
       ('User 73 logged in', 1962),
       ('Disk space running low: 30% remaining', 4066),
       ('Disk space running low: 48% remaining', 4111),
       ('Disk space running low: 5% remaining', 3967),
       ('Disk space running low: 31% remaining', 3984),
       ('User 78 logged in', 1955),
       ('Disk space running low: 12% remaining', 4041),
       ('User 80 logged in', 1992),
       ('Error in module moduleB: Array index out of range', 22245),
       ('Disk space running low: 28% remaining', 4077),
       ('User 14 logged in', 2026),
       ('Error in module moduleC: Segmentation fault', 22278),
       ('User 38 logged in', 2010),
       ('User 44 logged in', 2041),
       ('User 43 logged in', 2041),
       ('Disk space running low: 2% remaining', 3946),
       ('User 96 logged in', 1969),
       ('Disk space running low: 10% remaining', 3992),
       ('User 21 logged in', 1974),
       ('User 66 logged in', 1945),
       ('User 77 logged in', 2013),
       ('Disk space running low: 38% remaining', 3993),
       ('User 60 logged in', 2011),
       ('User 11 logged in', 1967),
       ('Disk space running low: 49% remaining', 3958),
       ('Disk space running low: 16% remaining', 4002),
       ('User 30 logged in', 1955),
       ('Disk space running low: 13% remaining', 3904),
       ('Disk space running low: 3% remaining', 3959),
       ('Disk space running low: 47% remaining', 3938),
       ('Disk space running low: 19% remaining', 3946),
       ('Disk space running low: 15% remaining', 4061),
       ('User 62 logged in', 2021),
       ('User 82 logged in', 1936),
       ('Disk space running low: 40% remaining', 4009),
       ('User 88 logged in', 2001),
       ('User 95 logged in', 1982),
       ('User 67 logged in', 2039),
       ('User 9 logged in', 1934),
       ('User 49 logged in', 1967),
       ('User 39 logged in', 1965),
       ('User 93 logged in', 2124),
       ('User 89 logged in', 1967),
       ('Disk space running low: 9% remaining', 3946),
       ('User 61 logged in', 2016),
       ('User 52 logged in', 2026),
       ('User 27 logged in', 2031),
       ('Disk space running low: 43% remaining', 4021),
       ('Disk space running low: 7% remaining', 4062),
       ('User 74 logged in', 2015),
       ('User 81 logged in', 2103),
       ('User 32 logged in', 1950),
       ('User 99 logged in', 1967),
       ('Disk space running low: 24% remaining', 3981),
       ('User 47 logged in', 2018),
       ('Disk space running low: 17% remaining', 3956),
       ('User 72 logged in', 1993),
       ('User 53 logged in', 2030),
       ('User 90 logged in', 2027),
       ('Disk space running low: 42% remaining', 3891),
       ('User 25 logged in', 2073),
       ('User 35 logged in', 1969),
       ('Disk space running low: 41% remaining', 3946),
       ('User 59 logged in', 2044),
       ('User 50 logged in', 2031),
       ('User 20 logged in', 1999),
       ('User 18 logged in', 1952),
       ('User 7 logged in', 1998),
       ('User 75 logged in', 2049),
       ('User 37 logged in', 2004),
       ('User 28 logged in', 2088),
       ('User 3 logged in', 2041),
       ('User 69 logged in', 2061),
       ('User 33 logged in', 1942),
       ('User 91 logged in', 1967),
       ('User 85 logged in', 1970),
       ('User 13 logged in', 1960),
       ('01 17:54:47', 124108),
       ('01 17:54:49', 119009),
       ('01 17:54:50', 121689),
       ('01 17:54:51', 119732),
       ('01 17:54:52', 120966),
       ('01 17:54:55', 23817),
       ('02', 1000000),
       ('01 17:54:46', 8211),
       ('CRITICAL', 199737),
       ('Connection timeout on server server3', 66475),
       ('Error in module moduleC: Null pointer exception', 22193),
       ('ERROR', 199434),
       ('Disk space running low: 46% remaining', 4003),
       ('Connection timeout on server server2', 66400),
       ('WARNING', 200727),
       ('User 68 logged in', 1979),
       ('Error in module moduleA: Null pointer exception', 22109),
       ("File 'file3.csv' uploaded successfully", 66889),
       ("File 'file2.txt' uploaded successfully", 66798),
       ('Error in module moduleA: Segmentation fault', 22175),
       ('Connection timeout on server server1', 67083),
       ('Disk space running low: 39% remaining', 3908),
       ('Disk space running low: 21% remaining', 3991),
       ('Error in module moduleB: Null pointer exception', 22051),
       ('Disk space running low: 18% remaining', 4114),
       ('Disk space running low: 14% remaining', 4036),
       ('Error in module moduleC: Array index out of range', 22260),
       ('User 1 logged in', 2040),
       ('Error in module moduleB: Segmentation fault', 22436),
       ('Error in module moduleA: Array index out of range', 22445),
       ('Disk space running low: 23% remaining', 4005),
       ('User 16 logged in', 1903),
       ('User 56 logged in', 1948),
       ('User 29 logged in', 2066),
       ('Disk space running low: 8% remaining', 3906),
       ('Disk space running low: 35% remaining', 4017),
       ('User 31 logged in', 2065),
       ('Disk space running low: 22% remaining', 3982),
       ('Disk space running low: 11% remaining', 3905),
       ('User 15 logged in', 1986),
       ('User 10 logged in', 1959),
       ('User 63 logged in', 1990),
       ('Disk space running low: 1% remaining', 4071),
       ('User 23 logged in', 1979),
       ('User 26 logged in', 2007),
       ('User 70 logged in', 1976),
       ('User 71 logged in', 1997),
       ('Disk space running low: 25% remaining', 4105),
       ('Disk space running low: 4% remaining', 3954),
       ('Disk space running low: 36% remaining', 4055),
       ('User 94 logged in', 1991),
       ('User 79 logged in', 2050),
       ('User 19 logged in', 2013),
       ('User 65 logged in', 1929),
       ('User 41 logged in', 1979),
       ('User 4 logged in', 1933),
       ('Disk space running low: 33% remaining', 4000),
       ('User 98 logged in', 1955),
       ('User 86 logged in', 2002),
       ('Disk space running low: 37% remaining', 4016),
       ('User 97 logged in', 1935),
       ('Disk space running low: 34% remaining', 4024),
       ('User 64 logged in', 2031),
       ('User 40 logged in', 1982),
       ('Disk space running low: 27% remaining', 4073),
       ('User 51 logged in', 1991),
       ('Disk space running low: 26% remaining', 3949),
       ('Disk space running low: 45% remaining', 4118),
       ('User 55 logged in', 1998),
       ('Disk space running low: 44% remaining', 4032),
       ('User 2 logged in', 2018),
       ('User 5 logged in', 2023),
       ('User 24 logged in', 1991),
       ('User 17 logged in', 1986),
       ('User 92 logged in', 1966),
       ('User 57 logged in', 2032),
       ('User 46 logged in', 1990),
       ('User 45 logged in', 1989),
       ('User 12 logged in', 1974),
       ('User 83 logged in', 2003),
       ('User 54 logged in', 2002),
       ('User 6 logged in', 1979),
       ('User 76 logged in', 1998),
       ('User 87 logged in', 1973),
       ('01 17:54:48', 123603),
       ('01 17:54:53', 118266),
       ('01 17:54:54', 120599)]
```

```
[16]: spark.stop()
```

```
[ ]:
```

0.1.1 Partition Demo

```
[18]: spark = SparkSession.builder \
              .appName("Partition Demo") \
              .getOrCreate()
```

```
25/02/10 16:45:40 INFO SparkEnv: Registering MapOutputTracker
25/02/10 16:45:40 INFO SparkEnv: Registering BlockManagerMaster
25/02/10 16:45:40 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
25/02/10 16:45:40 INFO SparkEnv: Registering OutputCommitCoordinator
```

```
[19]: spark
```

```
[19]: <pyspark.sql.session.SparkSession at 0x7fb1bc097190>
```

[21]:
```
log_rdd = spark.sparkContext.parallelize(hdfs_path)
```

[24]:
```
print(f"Default Partition {log_rdd.getNumPartitions()}")
```

```
Default Partition 2
```

Set partition to 200

[25]:
```
log_rdd = log_rdd.repartition(200)
```

[28]:
```
print(f"Partition {log_rdd.getNumPartitions()}")
```

```
Partition 200
```

[29]:
```
spark.stop()
```

[ ]:
```

```

[ ]:
```

```

0.1.2 RDD Operations

[30]:
```
spark = SparkSession.builder \
        .appName("RDD Operation") \
        .getOrCreate()
```

```
25/02/10 16:51:18 INFO SparkEnv: Registering MapOutputTracker
25/02/10 16:51:18 INFO SparkEnv: Registering BlockManagerMaster
25/02/10 16:51:18 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
25/02/10 16:51:18 INFO SparkEnv: Registering OutputCommitCoordinator
```

[32]:
```
spark
```

[32]:
```
<pyspark.sql.session.SparkSession at 0x7fb1bc258d30>
```

[33]:
```
!hadoop fs -ls /ecommerce_data/ecommerce_data/500MB
```

```
Found 5 items
-rw-r--r--   2 lokesh hadoop  570783961 2025-02-10 16:00
/ecommerce_data/ecommerce_data/500MB/customers.csv
-rw-r--r--   2 lokesh hadoop  480952071 2025-02-10 16:00
/ecommerce_data/ecommerce_data/500MB/items.csv
-rw-r--r--   2 lokesh hadoop  472632078 2025-02-10 16:00
/ecommerce_data/ecommerce_data/500MB/orders.csv
-rw-r--r--   2 lokesh hadoop  468231725 2025-02-10 16:01
/ecommerce_data/ecommerce_data/500MB/payments.csv
-rw-r--r--   2 lokesh hadoop  448185359 2025-02-10 16:01
/ecommerce_data/ecommerce_data/500MB/shippings.csv
```

[34]:
```
hdfs_path = "/ecommerce_data/ecommerce_data/500MB/customers.csv"
```

[37]:

```
customer_rdd = spark.sparkContext.textFile(hdfs_path)
```

[38]:

```
customer_rdd.take(5)
```

[38]:

```
['customer_id,name,city,state,country,registration_date,is_active',
 '0,Customer_0,Mumbai,Telangana,India,2023-03-21,True',
 '1,Customer_1,Chennai,West Bengal,India,2023-05-27,False',
 '2,Customer_2,Pune,Karnataka,India,2023-10-11,False',
 '3,Customer_3,Hyderabad,Gujarat,India,2023-11-11,False']
```

[39]:

```
header = customer_rdd.first()
```

[40]:

```
header
```

[40]:

```
'customer_id,name,city,state,country,registration_date,is_active'
```

[43]:

```
customer_filter_rdd = customer_rdd.filter(lambda row : row != header)

customer_filter_rdd.take(3)
```

[43]:

```
['0,Customer_0,Mumbai,Telangana,India,2023-03-21,True',
 '1,Customer_1,Chennai,West Bengal,India,2023-05-27,False',
 '2,Customer_2,Pune,Karnataka,India,2023-10-11,False']
```

0.1.3 Map

Parse data

[44]:

```
def parse_data(row):
    fields = row.split(",")

    return (
        int(fields[0]),
        fields[1],
        fields[2],
        fields[3],
        fields[4],
        fields[5],
        fields[6] == 'true'
    )
```

[47]:

```
customer_parsed_rdd = customer_filter_rdd.map(parse_data)

customer_parsed_rdd.take(5)
```

[47]:

```
[(0, 'Customer_0', 'Mumbai', 'Telangana', 'India', '2023-03-21', False),
 (1, 'Customer_1', 'Chennai', 'West Bengal', 'India', '2023-05-27', False),
 (2, 'Customer_2', 'Pune', 'Karnataka', 'India', '2023-10-11', False),
 (3, 'Customer_3', 'Hyderabad', 'Gujarat', 'India', '2023-11-11', False),
 (4, 'Customer_4', 'Mumbai', 'Karnataka', 'India', '2023-05-09', False)]
```

```
[48]: ## Reduce by key with map
      customer_parsed_rdd.map(lambda row : (row[2], 1)).reduceByKey(lambda x, y : x +␣
       ↪y).collect()
```

```
[48]: [('Kolkata', 1096777),
       ('Pune', 1095748),
       ('Chennai', 1095052),
       ('Hyderabad', 1096426),
       ('Ahmedabad', 1097162),
       ('Mumbai', 1095815),
       ('Delhi', 1096183),
       ('Bangalore', 1094195)]
```

```
[49]: spark.stop()
```

```
[ ]:
```

0.1.4 Wide V/S Narrow Transformation

```
[50]: spark = SparkSession.builder \
              .appName("Wide-Transformation") \
              .master('yarn') \
              .getOrCreate()
```

```
25/02/10 17:06:49 INFO SparkEnv: Registering MapOutputTracker
25/02/10 17:06:49 INFO SparkEnv: Registering BlockManagerMaster
25/02/10 17:06:49 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
25/02/10 17:06:49 INFO SparkEnv: Registering OutputCommitCoordinator
```

```
[52]: hdfs_path = "/ecommerce_data/ecommerce_data/10MB/customers.csv"
```

```
[53]: customer_rdd = spark.sparkContext.textFile(hdfs_path)
```

```
[55]: header = customer_rdd.first()

      customer_rdd = customer_rdd.filter(lambda row : row != header)
```

```
[57]: parsed_rdd = customer_rdd.map(parse_data)
```

```
[58]: active_customers = parsed_rdd.filter(lambda row:row[6])
```

```
[59]: grouped_by_city_rdd = parsed_rdd.map(lambda row: (row[2], 1)).
        ↪reduceByKey(lambda x, y: x + y)
```

```
[60]: grouped_by_city_rdd.collect()
```

```
[60]: [('Chennai', 21046),
       ('Mumbai', 21041),
       ('Pune', 21481),
       ('Bangalore', 21272),
       ('Hyderabad', 21174),
       ('Ahmedabad', 21272),
       ('Delhi', 21123),
       ('Kolkata', 21264)]
```

```
[61]: spark.stop()
```

```
[ ]:
```