

Caching - RDD

February 4, 2025

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

```
[24]: spark = SparkSession.builder.appName('Caching_Big_File').enableHiveSupport().  
      ↪getOrCreate()
```

```
25/02/02 04:55:05 INFO SparkEnv: Registering MapOutputTracker  
25/02/02 04:55:05 INFO SparkEnv: Registering BlockManagerMaster  
25/02/02 04:55:05 INFO SparkEnv: Registering BlockManagerMasterHeartbeat  
25/02/02 04:55:05 INFO SparkEnv: Registering OutputCommitCoordinator
```

```
[25]: spark
```

```
[25]: <pyspark.sql.session.SparkSession at 0x7f42c7484070>
```

initially you said we can not control caching , while in this 3rd point you mention to avoid caching in case of memory pressure? If we create multiple data frames in a spark application, won't they be stored in memory ? why do we need to cache them again ?

If memory is insufficient in caching, how does the data stored in disk? Is it serialized or deserialized?

```
[26]: customers_rdd = spark.sparkContext.textFile('/tmp/customers_500mb.csv')
```

```
[27]: customers_filtered = customers_rdd.filter(lambda row : 'Mumbai' in row)  
      customers_mapped = customers_filtered.map(lambda row : (row.split(',')[0],1))  
      customers_reduced = customers_mapped.reduceByKey(lambda x,y:x+y)
```

```
[28]: customers_reduced.count()
```

```
[28]: 1095815
```

```
[31]: customers_reduced.count()
```

```
[31]: 1095815
```

```
[32]: customers_reduced.cache()
```

```
[32]: PythonRDD[10] at RDD at PythonRDD.scala:53
```

```
[35]: customers_reduced.count()
```

```
[35]: 1095815
```

```
[36]: customers_reduced.count()
```

```
[36]: 1095815
```

```
[ ]:
```

```
[ ]: customers_reduced.count()
```

```
[37]: customers_reduced.collect()
```

```
[37]: [('42', 1),  
      ('47', 1),  
      ('61', 1),  
      ('129', 1),  
      ('219', 1),  
      ('261', 1),  
      ('299', 1),  
      ('326', 1),  
      ('367', 1),  
      ('464', 1),  
      ('465', 1),  
      ('496', 1),  
      ('549', 1),  
      ('556', 1),  
      ('559', 1),  
      ('584', 1),  
      ('637', 1),  
      ('660', 1),  
      ('663', 1),  
      ('686', 1),  
      ('752', 1),  
      ('753', 1),  
      ('763', 1),  
      ('772', 1),  
      ('933', 1),  
      ('935', 1),  
      ('960', 1),  
      ('1023', 1),  
      ('1103', 1),  
      ('1104', 1),  
      ('1147', 1),  
      ('1180', 1),  
      ('1200', 1),  
      ('1386', 1),
```

('1394', 1),
('1414', 1),
('1461', 1),
('1504', 1),
('1533', 1),
('1642', 1),
('1712', 1),
('1731', 1),
('1732', 1),
('1742', 1),
('1750', 1),
('1763', 1),
('1764', 1),
('1768', 1),
('1815', 1),
('1843', 1),
('1871', 1),
('1940', 1),
('1980', 1),
('1984', 1),
('1987', 1),
('1993', 1),
('2006', 1),
('2096', 1),
('2099', 1),
('2119', 1),
('2185', 1),
('2215', 1),
('2253', 1),
('2263', 1),
('2363', 1),
('2412', 1),
('2437', 1),
('2447', 1),
('2467', 1),
('2494', 1),
('2667', 1),
('2696', 1),
('2713', 1),
('2812', 1),
('2862', 1),
('2865', 1),
('2921', 1),
('2923', 1),
('2929', 1),
('3032', 1),
('3155', 1),

('3162', 1),
('3216', 1),
('3225', 1),
('3245', 1),
('3258', 1),
('3267', 1),
('3291', 1),
('3353', 1),
('3412', 1),
('3459', 1),
('3461', 1),
('3473', 1),
('3522', 1),
('3562', 1),
('3627', 1),
('3634', 1),
('3644', 1),
('3687', 1),
('3713', 1),
('3717', 1),
('3776', 1),
('3821', 1),
('3905', 1),
('3910', 1),
('3937', 1),
('3974', 1),
('4058', 1),
('4137', 1),
('4154', 1),
('4193', 1),
('4212', 1),
('4287', 1),
('4289', 1),
('4425', 1),
('4448', 1),
('4459', 1),
('4508', 1),
('4511', 1),
('4520', 1),
('4554', 1),
('4614', 1),
('4779', 1),
('4783', 1),
('4795', 1),
('4798', 1),
('4814', 1),
('4841', 1),

('4995', 1),
('5022', 1),
('5054', 1),
('5084', 1),
('5165', 1),
('5184', 1),
('5209', 1),
('5254', 1),
('5262', 1),
('5267', 1),
('5335', 1),
('5379', 1),
('5399', 1),
('5421', 1),
('5422', 1),
('5440', 1),
('5446', 1),
('5495', 1),
('5582', 1),
('5590', 1),
('5594', 1),
('5625', 1),
('5662', 1),
('5672', 1),
('5704', 1),
('5751', 1),
('5756', 1),
('5848', 1),
('5854', 1),
('6013', 1),
('6027', 1),
('6089', 1),
('6133', 1),
('6201', 1),
('6219', 1),
('6245', 1),
('6317', 1),
('6326', 1),
('6340', 1),
('6355', 1),
('6369', 1),
('6374', 1),
('6394', 1),
('6408', 1),
('6451', 1),
('6464', 1),
('6488', 1),

('6537', 1),
('6553', 1),
('6559', 1),
('6603', 1),
('6637', 1),
('6656', 1),
('6658', 1),
('6676', 1),
('6730', 1),
('6738', 1),
('6826', 1),
('6839', 1),
('6898', 1),
('7025', 1),
('7043', 1),
('7109', 1),
('7124', 1),
('7157', 1),
('7180', 1),
('7209', 1),
('7215', 1),
('7218', 1),
('7305', 1),
('7357', 1),
('7381', 1),
('7405', 1),
('7409', 1),
('7459', 1),
('7476', 1),
('7543', 1),
('7586', 1),
('7588', 1),
('7630', 1),
('7641', 1),
('7663', 1),
('7678', 1),
('7700', 1),
('7784', 1),
('7795', 1),
('7823', 1),
('7848', 1),
('7916', 1),
('8010', 1),
('8015', 1),
('8038', 1),
('8050', 1),
('8059', 1),

('8081', 1),
('8085', 1),
('8161', 1),
('8179', 1),
('8226', 1),
('8262', 1),
('8293', 1),
('8392', 1),
('8413', 1),
('8454', 1),
('8458', 1),
('8501', 1),
('8533', 1),
('8534', 1),
('8550', 1),
('8581', 1),
('8587', 1),
('8597', 1),
('8617', 1),
('8623', 1),
('8694', 1),
('8712', 1),
('8730', 1),
('8863', 1),
('8875', 1),
('8941', 1),
('9055', 1),
('9097', 1),
('9162', 1),
('9193', 1),
('9200', 1),
('9230', 1),
('9233', 1),
('9269', 1),
('9286', 1),
('9345', 1),
('9460', 1),
('9508', 1),
('9513', 1),
('9517', 1),
('9522', 1),
('9555', 1),
('9565', 1),
('9612', 1),
('9618', 1),
('9634', 1),
('9702', 1),

('9777', 1),
('9883', 1),
('9943', 1),
('9990', 1),
('9992', 1),
('10027', 1),
('10152', 1),
('10167', 1),
('10202', 1),
('10206', 1),
('10225', 1),
('10226', 1),
('10296', 1),
('10297', 1),
('10373', 1),
('10392', 1),
('10418', 1),
('10422', 1),
('10428', 1),
('10466', 1),
('10488', 1),
('10549', 1),
('10556', 1),
('10582', 1),
('10608', 1),
('10739', 1),
('10746', 1),
('10772', 1),
('10797', 1),
('10798', 1),
('10823', 1),
('10885', 1),
('10907', 1),
('10915', 1),
('10929', 1),
('10949', 1),
('10970', 1),
('10989', 1),
('11000', 1),
('11047', 1),
('11066', 1),
('11088', 1),
('11103', 1),
('11128', 1),
('11140', 1),
('11151', 1),
('11186', 1),

('11245', 1),
('11323', 1),
('11366', 1),
('11377', 1),
('11394', 1),
('11406', 1),
('11420', 1),
('11421', 1),
('11423', 1),
('11517', 1),
('11563', 1),
('11593', 1),
('11631', 1),
('11652', 1),
('11690', 1),
('11702', 1),
('11704', 1),
('11741', 1),
('11774', 1),
('11779', 1),
('11810', 1),
('11844', 1),
('11851', 1),
('11890', 1),
('11914', 1),
('11942', 1),
('11991', 1),
('12037', 1),
('12043', 1),
('12132', 1),
('12141', 1),
('12142', 1),
('12176', 1),
('12205', 1),
('12226', 1),
('12229', 1),
('12290', 1),
('12315', 1),
('12332', 1),
('12356', 1),
('12357', 1),
('12379', 1),
('12415', 1),
('12438', 1),
('12453', 1),
('12465', 1),
('12502', 1),

('12516', 1),
('12575', 1),
('12603', 1),
('12652', 1),
('12679', 1),
('12713', 1),
('12747', 1),
('12751', 1),
('12887', 1),
('12906', 1),
('12946', 1),
('12972', 1),
('12979', 1),
('13035', 1),
('13117', 1),
('13156', 1),
('13195', 1),
('13237', 1),
('13478', 1),
('13530', 1),
('13570', 1),
('13599', 1),
('13606', 1),
('13772', 1),
('13789', 1),
('13867', 1),
('13877', 1),
('13900', 1),
('13913', 1),
('13926', 1),
('13939', 1),
('13958', 1),
('13969', 1),
('14031', 1),
('14035', 1),
('14048', 1),
('14056', 1),
('14093', 1),
('14102', 1),
('14129', 1),
('14209', 1),
('14224', 1),
('14227', 1),
('14234', 1),
('14322', 1),
('14373', 1),
('14396', 1),

('14442', 1),
('14443', 1),
('14501', 1),
('14526', 1),
('14607', 1),
('14639', 1),
('14719', 1),
('14734', 1),
('14829', 1),
('14866', 1),
('14895', 1),
('14952', 1),
('14975', 1),
('15028', 1),
('15051', 1),
('15116', 1),
('15145', 1),
('15237', 1),
('15248', 1),
('15263', 1),
('15458', 1),
('15459', 1),
('15491', 1),
('15499', 1),
('15518', 1),
('15617', 1),
('15618', 1),
('15636', 1),
('15641', 1),
('15667', 1),
('15765', 1),
('15790', 1),
('15800', 1),
('15828', 1),
('15833', 1),
('15838', 1),
('15839', 1),
('15849', 1),
('15886', 1),
('15889', 1),
('15969', 1),
('16009', 1),
('16090', 1),
('16121', 1),
('16125', 1),
('16164', 1),
('16165', 1),

('16170', 1),
('16213', 1),
('16224', 1),
('16288', 1),
('16356', 1),
('16366', 1),
('16408', 1),
('16543', 1),
('16607', 1),
('16631', 1),
('16638', 1),
('16649', 1),
('16670', 1),
('16728', 1),
('16740', 1),
('16747', 1),
('16774', 1),
('16839', 1),
('16870', 1),
('16889', 1),
('16896', 1),
('16911', 1),
('16939', 1),
('16952', 1),
('17077', 1),
('17084', 1),
('17086', 1),
('17128', 1),
('17136', 1),
('17138', 1),
('17186', 1),
('17250', 1),
('17266', 1),
('17270', 1),
('17277', 1),
('17366', 1),
('17413', 1),
('17415', 1),
('17450', 1),
('17475', 1),
('17483', 1),
('17484', 1),
('17561', 1),
('17563', 1),
('17598', 1),
('17684', 1),
('17764', 1),

('17774', 1),
('17789', 1),
('17807', 1),
('17810', 1),
('17812', 1),
('17842', 1),
('17845', 1),
('17853', 1),
('17863', 1),
('17878', 1),
('17898', 1),
('17928', 1),
('17949', 1),
('17968', 1),
('18054', 1),
('18074', 1),
('18089', 1),
('18094', 1),
('18098', 1),
('18110', 1),
('18347', 1),
('18429', 1),
('18511', 1),
('18574', 1),
('18649', 1),
('18656', 1),
('18670', 1),
('18714', 1),
('18792', 1),
('18799', 1),
('18804', 1),
('18927', 1),
('19091', 1),
('19130', 1),
('19144', 1),
('19264', 1),
('19292', 1),
('19376', 1),
('19385', 1),
('19476', 1),
('19504', 1),
('19605', 1),
('19664', 1),
('19689', 1),
('19728', 1),
('19744', 1),
('19794', 1),

('19855', 1),
('20003', 1),
('20005', 1),
('20023', 1),
('20076', 1),
('20140', 1),
('20197', 1),
('20212', 1),
('20256', 1),
('20259', 1),
('20281', 1),
('20341', 1),
('20363', 1),
('20390', 1),
('20482', 1),
('20563', 1),
('20577', 1),
('20615', 1),
('20634', 1),
('20643', 1),
('20685', 1),
('20695', 1),
('20722', 1),
('20743', 1),
('20814', 1),
('20821', 1),
('20870', 1),
('20976', 1),
('21010', 1),
('21034', 1),
('21075', 1),
('21172', 1),
('21182', 1),
('21191', 1),
('21234', 1),
('21427', 1),
('21442', 1),
('21452', 1),
('21463', 1),
('21521', 1),
('21600', 1),
('21601', 1),
('21608', 1),
('21642', 1),
('21651', 1),
('21678', 1),
('21769', 1),

('21945', 1),
('21969', 1),
('21977', 1),
('21978', 1),
('22025', 1),
('22040', 1),
('22042', 1),
('22085', 1),
('22224', 1),
('22322', 1),
('22331', 1),
('22332', 1),
('22353', 1),
('22365', 1),
('22458', 1),
('22465', 1),
('22504', 1),
('22528', 1),
('22573', 1),
('22579', 1),
('22582', 1),
('22601', 1),
('22604', 1),
('22630', 1),
('22692', 1),
('22759', 1),
('22760', 1),
('22782', 1),
('22790', 1),
('22937', 1),
('22974', 1),
('22998', 1),
('23029', 1),
('23055', 1),
('23071', 1),
('23079', 1),
('23113', 1),
('23125', 1),
('23159', 1),
('23261', 1),
('23290', 1),
('23293', 1),
('23316', 1),
('23373', 1),
('23495', 1),
('23572', 1),
('23592', 1),

('23596', 1),
('23631', 1),
('23645', 1),
('23654', 1),
('23816', 1),
('23903', 1),
('23908', 1),
('23928', 1),
('23932', 1),
('23983', 1),
('23988', 1),
('24091', 1),
('24092', 1),
('24188', 1),
('24197', 1),
('24199', 1),
('24206', 1),
('24256', 1),
('24427', 1),
('24520', 1),
('24527', 1),
('24531', 1),
('24538', 1),
('24543', 1),
('24557', 1),
('24583', 1),
('24618', 1),
('24650', 1),
('24666', 1),
('24704', 1),
('24721', 1),
('24749', 1),
('24750', 1),
('24877', 1),
('24982', 1),
('25021', 1),
('25133', 1),
('25137', 1),
('25140', 1),
('25157', 1),
('25186', 1),
('25238', 1),
('25331', 1),
('25335', 1),
('25348', 1),
('25352', 1),
('25429', 1),

('25491', 1),
('25499', 1),
('25502', 1),
('25531', 1),
('25651', 1),
('25654', 1),
('25658', 1),
('25699', 1),
('25757', 1),
('25767', 1),
('25807', 1),
('25871', 1),
('25982', 1),
('26039', 1),
('26049', 1),
('26057', 1),
('26078', 1),
('26080', 1),
('26091', 1),
('26100', 1),
('26111', 1),
('26136', 1),
('26187', 1),
('26245', 1),
('26270', 1),
('26341', 1),
('26408', 1),
('26478', 1),
('26493', 1),
('26494', 1),
('26595', 1),
('26630', 1),
('26645', 1),
('26647', 1),
('26651', 1),
('26652', 1),
('26695', 1),
('26710', 1),
('26737', 1),
('26764', 1),
('26823', 1),
('26827', 1),
('26828', 1),
('26851', 1),
('26908', 1),
('26932', 1),
('26935', 1),

('27023', 1),
('27049', 1),
('27058', 1),
('27063', 1),
('27072', 1),
('27147', 1),
('27196', 1),
('27285', 1),
('27326', 1),
('27395', 1),
('27479', 1),
('27517', 1),
('27633', 1),
('27675', 1),
('27683', 1),
('27739', 1),
('27796', 1),
('27872', 1),
('27904', 1),
('27923', 1),
('28035', 1),
('28061', 1),
('28099', 1),
('28119', 1),
('28137', 1),
('28158', 1),
('28169', 1),
('28280', 1),
('28295', 1),
('28341', 1),
('28374', 1),
('28381', 1),
('28455', 1),
('28462', 1),
('28486', 1),
('28627', 1),
('28758', 1),
('28764', 1),
('28765', 1),
('28872', 1),
('28918', 1),
('28974', 1),
('29026', 1),
('29074', 1),
('29081', 1),
('29093', 1),
('29100', 1),

('29133', 1),
('29215', 1),
('29221', 1),
('29234', 1),
('29256', 1),
('29330', 1),
('29390', 1),
('29435', 1),
('29446', 1),
('29463', 1),
('29475', 1),
('29519', 1),
('29666', 1),
('29687', 1),
('29730', 1),
('29752', 1),
('29851', 1),
('29943', 1),
('29999', 1),
('30033', 1),
('30043', 1),
('30079', 1),
('30108', 1),
('30113', 1),
('30174', 1),
('30180', 1),
('30215', 1),
('30249', 1),
('30273', 1),
('30291', 1),
('30294', 1),
('30300', 1),
('30342', 1),
('30400', 1),
('30410', 1),
('30484', 1),
('30502', 1),
('30524', 1),
('30542', 1),
('30573', 1),
('30611', 1),
('30628', 1),
('30641', 1),
('30675', 1),
('30736', 1),
('30738', 1),
('30770', 1),

('30781', 1),
('30811', 1),
('30813', 1),
('30930', 1),
('30956', 1),
('31040', 1),
('31093', 1),
('31154', 1),
('31211', 1),
('31228', 1),
('31254', 1),
('31340', 1),
('31360', 1),
('31391', 1),
('31399', 1),
('31497', 1),
('31507', 1),
('31557', 1),
('31575', 1),
('31576', 1),
('31612', 1),
('31631', 1),
('31665', 1),
('31713', 1),
('31715', 1),
('31745', 1),
('31821', 1),
('31901', 1),
('31954', 1),
('31980', 1),
('31988', 1),
('32173', 1),
('32224', 1),
('32250', 1),
('32279', 1),
('32312', 1),
('32338', 1),
('32373', 1),
('32376', 1),
('32402', 1),
('32477', 1),
('32483', 1),
('32515', 1),
('32520', 1),
('32612', 1),
('32648', 1),
('32680', 1),

('32760', 1),
('32763', 1),
('32773', 1),
('32785', 1),
('32788', 1),
('32866', 1),
('32924', 1),
('32933', 1),
('32935', 1),
('32938', 1),
('32946', 1),
('32964', 1),
('33095', 1),
('33147', 1),
('33163', 1),
('33256', 1),
('33257', 1),
('33354', 1),
('33423', 1),
('33425', 1),
('33484', 1),
('33494', 1),
('33509', 1),
('33559', 1),
('33582', 1),
('33587', 1),
('33678', 1),
('33908', 1),
('33915', 1),
('33924', 1),
('33982', 1),
('33988', 1),
('34008', 1),
('34036', 1),
('34049', 1),
('34072', 1),
('34078', 1),
('34092', 1),
('34093', 1),
('34155', 1),
('34240', 1),
('34402', 1),
('34442', 1),
('34662', 1),
('34677', 1),
('34706', 1),
('34713', 1),

('34715', 1),
('34717', 1),
('34743', 1),
('34744', 1),
('34766', 1),
('34847', 1),
('34923', 1),
('34961', 1),
('34964', 1),
('34985', 1),
('34990', 1),
('35064', 1),
('35133', 1),
('35148', 1),
('35178', 1),
('35229', 1),
('35244', 1),
('35276', 1),
('35329', 1),
('35363', 1),
('35376', 1),
('35427', 1),
('35443', 1),
('35513', 1),
('35557', 1),
('35641', 1),
('35650', 1),
('35676', 1),
('35718', 1),
('35764', 1),
('35772', 1),
('35779', 1),
('35825', 1),
('35842', 1),
('35876', 1),
('35909', 1),
('36036', 1),
('36114', 1),
('36133', 1),
('36135', 1),
('36157', 1),
('36190', 1),
('36254', 1),
('36281', 1),
('36301', 1),
('36331', 1),
('36385', 1),

```
('36391', 1),
('36457', 1),
('36485', 1),
('36523', 1),
('36531', 1),
('36587', 1),
('36602', 1),
('36616', 1),
('36641', 1),
('36672', 1),
('36692', 1),
('36702', 1),
('36722', 1),
('36733', 1),
('36791', 1),
('36819', 1),
('36856', 1),
('36868', 1),
('36876', 1),
('36906', 1),
('36915', 1),
('36932', 1),
('36947', 1),
('36956', 1),
('37042', 1),
('37049', 1),
...]
```

```
[38]: customers_final = customers_reduced.filter(lambda row: int(row[0])<500)
```

```
[ ]: customers_final.collect()
```

```
[40]: customers_reduced.unpersist()
```

```
[40]: PythonRDD[10] at RDD at PythonRDD.scala:53
```

```
[ ]:
```

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