



***Name: Drishti Harshit Doshi***

***Student ID: 2203898***

***Module Code: CS5810***

***Module Title: High-Performance Computational Infrastructures***

***Academic Year: 2022-2023***

***Assignment Title: Coursework Assignment using MapReduce***

## Section 1:

### Introduction:

**Big Data** refers to the collection of datasets that are large and complex and which cannot be stored or processed using traditional data processing systems. Organizations are realizing the importance of big data analytics for getting useful business insights. They are using large datasets to explore hidden patterns, find correlations, understand customer preferences, and other useful information. Analyzing big data helps the organization generate more revenue, formulate effective marketing strategies, provide better customer service by preventing customer churn, and compete well with its opponents.

The traditional approach of using a **Relational Database Management system** has certain **limitations** which affect the business of the organization. Some of them include:

- Storing a large amount of data generated by companies every second.
- Storing heterogeneous data from the website, API, social media, text data, etc.
- Accessing and processing speed to handle big data is less.

In the year 2008, **Yahoo released Hadoop** as an open-source project to Apache Software Foundation.

Hadoop is a framework for storing Big-Data sets in a distributed environment that can be processed parallelly.

**Hadoop** has the following main components:

- Hadoop Distributed File System (HDFS): A block-structured file system where each file is divided into blocks of a pre-determined size. These blocks are stored across a cluster of one or several machines. HDFS runs on standard or low-end hardware. It has high fault tolerance and handles large datasets effectively. HDFS is preferable for low-latency access.
- Yet Another Resource Negotiator (YARN): It is a Hadoop cluster resource management system. It Performs all processing activities by allocating resources and scheduling tasks.
- Map-reduce: It is a programming model for data processing. MapReduce has a framework that allows the programs to be executed on parallel computers. The map task (mapper) takes input data and converts it into a dataset computed in the form of key-value pairs. The reduce tasks (reducer) aggregates and provides the output.

### Problem Description:

Every day, supermarkets sell millions of products to millions of people. In developed countries, people are dependent on supermarkets for their daily necessities. In addition to product pricing, providing good customer service and allowances are additional factors for competing with other stores. During this period, the retail sector is heavily facing crises due to global economic recession and financial instability. With the rise in technology and users, it is becoming increasingly difficult for retailers to meet customers' expectations. E-commerce, entrepreneurship, innovations, increase in private businesses, etc., all these factors have caused a massive increase in competition.

**Use case:** The US retail store aims to improve customer service as a strategy to prevent customer churn and retain existing customers. Big-data technology can be used here to extract insights into customer details that contribute to higher sales of the store and lower sales of the store. Accordingly, the Supermarket stores will focus on the strategies that will increase customer loyalty i.e., by providing a personalized experience, and promotions that will help to achieve customer retention.

### Research Question:

What is the total amount spent by each customer on their purchases, combined?

### Dataset:

The US Superstore Sales data has been used for the analysis. The dataset has been collected from Kaggle. The dataset has two files in .xls format:

- a) **Customer data:** Data of the customers buying the products.

Name	Description	Datatype
Customer ID	A unique ID assigned to each customer	String
Customer Name	Name of customers	String
Segment	Consumer or Co-operate sector	String
Country	Country of the Superstore outlet	String
City	City of the Superstore outlet	String
State	State of the Superstore outlet	String
Postal Code	Postal code of the Superstore outlet	String

- b) **Transaction data:** Transaction details of the superstore.

Name	Description	Datatype
Transaction ID	Transaction details of orders	String
Order Date	Date of transaction	Ordinal
Customer ID	A unique ID assigned to each customer	String
Sales	Sales of each transaction	Numerical

### **Desired Output:**

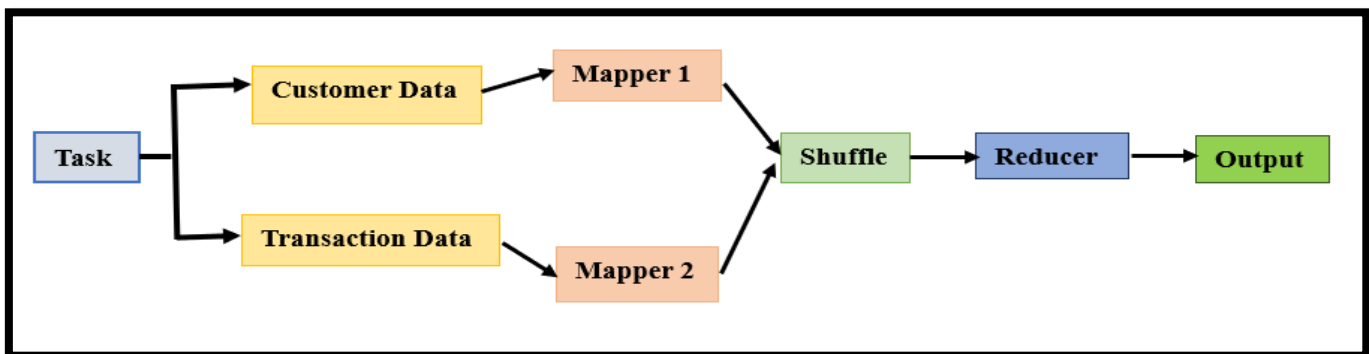
Customer Name	Total number of transactions	Total Sales

The Hadoop MapReduce framework can perform Joins by combining two or more database tables on foreign keys. Customer ID is the common variable in both data files. **The Customer ID** (foreign key) can be used to **combine** both datasets and get the total number of transactions and total sales for each customer.

There are two types of Joins in MapReduce:

- Map-side Join: The join operation is performed by the mapper. It is convenient when one of the tables is small (Distributed cache) and not recommended for larger tables.
- Reduce-side Join: The join operation is performed by the reducer. They are simpler and most widely used joins because Hadoop sends identical keys to the same reducer.

The given research question: What is the total amount spent by each customer on their purchases, combined? will be answered using **Reduce-Side Joins** of Hadoop MapReduce.



## Section 2:

### Design & Implementation:

Reduce-side joins have been implemented to find the research answer using MapReduce in the Hadoop framework. The research question asks to find the total amount of purchases made by each customer, by combining their multiple transactions and assigning the total number of sales, and the total number of transactions for each customer as the final output. Firstly, using the Virtual Machine which is provided by the Computer Science department at Brunel University London. Using Eclipse IDE for writing the JAVA MapReduce code. Steps for opening the JAVA project: Open Eclipse, open a new JAVA project, export external libraries, and create four classes in the given order- CustomerMapperfile, TransactionMapperfile, ReducerFile, and DriverFile using package: *org.apache.hadoop.mapreduce.join*

#### 1) Customer Mapper:

##### *Customer Mapperfile:*

```
*CustomerMapperfile.java  TransactionMapperfile.java  ReducerFile.java  DriverFile.java

1 package org.apache.hadoop.mapreduce.join;
2
3 //Importing the libraries that contain the classes and methods needed in the Mapper class//
4 import java.io.IOException;
5 import org.apache.hadoop.io.LongWritable;
6 import org.apache.hadoop.io.Text;
7 import org.apache.hadoop.mapreduce.Mapper;
8
```

workspace - Mapreducecoursework/src/org/apache/hadoop/mapreduce/join/CustomerMapperfile.java - Eclipse IDE

Edit Source Refactor Navigate Search Project Run Window Help

Input Key and Value datatypes

Output Key and Value datatypes

The context object created here from the Context class allows the Mapper/Reducer to interact with the rest of the Hadoop system.

```
*CustomerMapperfile.java  TransactionMapperfile.java  ReducerFile.java  DriverFile.java

8
9 //Creating a subclass CustomerMapperFile using the keyword extends //
10 public class CustomerMapperfile extends Mapper<LongWritable, Text, Text, Text>
11 {
12
13 /*In order to handle the Objects in Hadoop way, Hadoop uses Text instead of java's String.
14 The Text class in Hadoop is similar to a java String, however, Text implements interfaces
15 like Comparable, Writable and WritableComparable.*/
16
17 // The map method takes three parameters//
18 public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException
19 {
20     String record = value.toString(); //create a variable record of a data type String
21                                     //convert value into string and store it in record
22
23     String[] parts = record.split("\t"); // splitting string record into an array of strings using the tab character ("\t")
24
25     context.write(new Text(parts[0]), new Text("CustomerData\t"+parts[1]));
26 // Emitting first part of the Customer file ie CustomerID (key)
27 // string "CustomerData" is a (value) concatenated with the second element ie Customer Name separated by a tab character//
28
29 }
30 }
```

The Customer Mapperfile: Importing necessary libraries. Creating a new class named CustomerMapperfile. Using **extend** to assign data types. The **input key** is **LongWritable**, **input value** is **Text**. The **output key** is **Text** and the **output value** is **Text**. One tuple is taken as an input at a time i.e., one input = one tuple. Then each word in that tuple is tokenized. The Customer ID along with Customer Name is retrieved. The Customer ID will be the Mapper Key. The tag "CustomerData" is added to assign each input tuple to CustomerMapperfile.

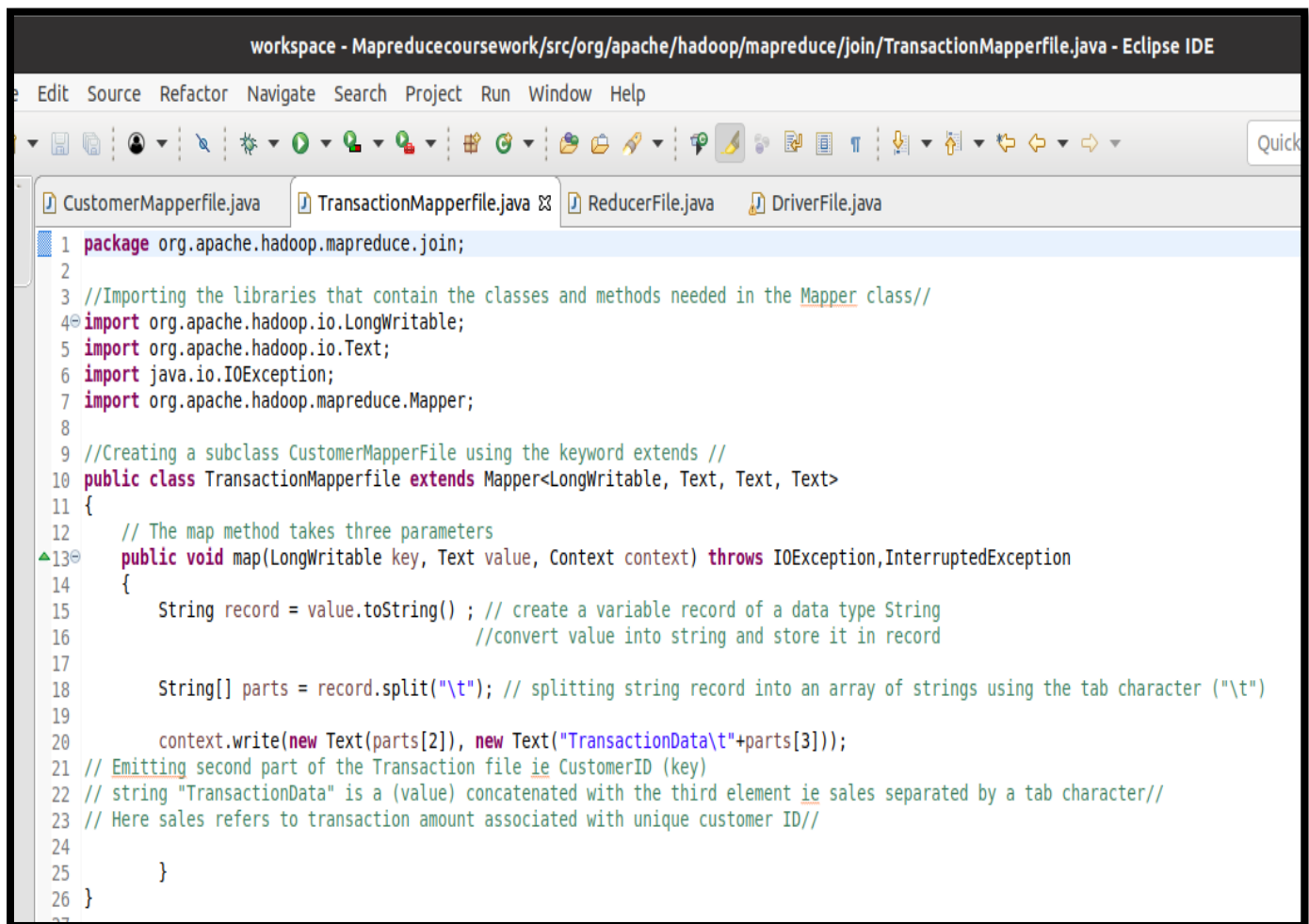
End output of CustomerMapperfile: Key-Value Pair: [Customer ID, Customer Name]

## 2) Transaction Mapper:

It will have same steps above, but the only difference will be in the fetching process. Sales will be taken instead of Customer Name and TransactionData will be taken as a tag to assign each input tuple to TransactionMapperfile.

End output of Transaction Mapper: [Customer ID, Sales]

### Transaction Mapperfile:



```

workspace - Mapreducecoursework/src/org/apache/hadoop/mapreduce/Join/TransactionMapperfile.java - Eclipse IDE
File Edit Source Refactor Navigate Search Project Run Window Help
CustomerMapperfile.java TransactionMapperfile.java ReducerFile.java DriverFile.java
1 package org.apache.hadoop.mapreduce.join;
2
3 //Importing the libraries that contain the classes and methods needed in the Mapper class//
4 import org.apache.hadoop.io.LongWritable;
5 import org.apache.hadoop.io.Text;
6 import java.io.IOException;
7 import org.apache.hadoop.mapreduce.Mapper;
8
9 //Creating a subclass CustomerMapperFile using the keyword extends //
10 public class TransactionMapperfile extends Mapper<LongWritable, Text, Text, Text>
11 {
12     // The map method takes three parameters
13     public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException
14     {
15         String record = value.toString(); // create a variable record of a data type String
16         //convert value into string and store it in record
17
18         String[] parts = record.split("\t"); // splitting string record into an array of strings using the tab character ("\t")
19
20         context.write(new Text(parts[2]), new Text("TransactionData\t"+parts[3]));
21 // Emitting second part of the Transaction file ie CustomerID (key)
22 // string "TransactionData" is a (value) concatenated with the third element ie sales separated by a tab character//
23 // Here sales refers to transaction amount associated with unique customer ID//
24     }
25 }
26

```

## 3) Sorting and Shuffling Phase:

The Sorting and Shuffling Phase will generate **an array list of values** assigned to **each unique key**.

The output at the end:

{Customer ID 1 – [(Customer Name 1), (Sales 1) , (Sales 2) , (Sales 3) , (Sales 4),...]}

{Customer ID 2 – [(Customer Name 2), (Sales 1) , (Sales 2) , (Sales 3) , (Sales 4),...]}

**Addition of  
sales at every  
transaction**

## ReducerFile:

The screenshot shows the ReducerFile.java code in an IDE. Annotations with arrows point to specific parts of the code:

- Input Key and Value Datatypes**: Points to the generic parameters `Text` in `Reducer<Text, Text, Text, Text>`.
- Out Key and Value Datatypes**: Points to the `Text` parameter in the `reduce` method signature.
- collection of Text objects that can be iterated over.**: Points to the `Iterable<Text> values` parameter in the `reduce` method.

```
1 package org.apache.hadoop.mapreduce.join;
2
3 //Importing the libraries that contain the classes and methods needed
4 import org.apache.hadoop.io.Text;
5 import java.io.IOException;
6 import org.apache.hadoop.mapreduce.Reducer;
7
8 /*Creating a subclass called ReducerFile from a parent
9 class called Reducer using the keyword extends */
10 public class ReducerFile extends
11     Reducer<Text, Text, Text, Text> {
12     // the reduce method takes three parameters
13     public void reduce (Text key, Iterable<Text> values, Context context)
14         throws IOException, InterruptedException {
15         String name = ""; // String variable named as name with empty string
16         double total = 0.0; // Double variable named as total with value of 0.0
17         int count = 0; // Integer variable named count with value of 0
18
19
20
21
22
23         for (Text t : values) { // Iterating through Text object ie values from 2 Mappers
24
25             String parts[] = t.toString().split("\t"); /*splitting each Text object into a String array
26                                                         using the tab character as the delimiter*/
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
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
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2
```

#### 4) Reducer Phase:

The framework will call reduce method() for each unique join key ie Customer ID and list of values ie Customer Name and an array of Sales.

Reduce Method: (reduce(Text key, Iterable<Text>values, Context context)

Reducer will perform **join operation** on the values present in the respective list of values. This will give the desired output eventually.

Each reducer task will be performed for each key ie Customer ID => Number of reducer = Number of Customer ID

Each reducer will have: key (unique customer ID) and value(Customer Name from CustomerMapperfile and Sales from TransactionMapperfile)

This will be followed by looping over all the values present in the list of reducer. Using if-elseif statement

The list of values will be splitted to check if it belongs to CustomerData or TransactionData.

If it belongs to TransactionData: then the counter value will be increased by one to calculate the frequency of transactions by that particular customer. This will be cumulatively added through each iterations. In the end, total amount of sales will be retrieved after summation of sales at every transaction.

Alternatively, if the value is from the CustomerData, then it will be stored in a string variable "name"

At the end of reduce phase, customer names = key

number of transactions and total sales = value

#### 5) Driver Phase:

Importing multiple input packages.Creating a user-defined job named Reduce-join.

Specifying the driver and reducer class in the Jar File

Assigning datatype of output key as Text and output value as Text which corresponds to String in JAVA.

Setting multiple input path to assign:

*For input path 1:*

Customer Name: passing argument [1]

Input format = Text Input Format class

Mapper = CustomerMapper class

*For input path 2:*

Number of transactions , sales :passing argument [2]

Input format = Text Input Format class

Mapper = TransactionMapper class

Assigning argument[3] for setting output path where the final output of Mapreduce file will be stored.

#### HDFS and Linux:

This will be followed by exporting the Jar file **ie HPC.jar** along with its specified class. Then open the Linux command terminal on the Hadoop environment. Formatting the name node. This will be followed by starting the hdfs and yarn system. Checking if all the six nodes are working using jps and the local host server. If all the six nodes ie name node, data node, resource manager, secondary name node, node manager, and jps working.

Creating a new directory in the HDFS named /input\_file. Download two datafiles named CustomerData.txt and TransactionData.txt and put them in the HDFS directory /input\_file. Execute the MapReduce program HPC.jar using Hadoop jar and store the output in the folder named output.



## DriverFile:

```
CustomerMapperfile.java  TransactionMapperfile.java  ReducerFile.java  DriverFile.java  ⌕
1 package org.apache.hadoop.mapreduce.join;
2 //Importing the libraries that contain the classes and methods needed in the driver class
3 import java.io.IOException;
4 import org.apache.hadoop.conf.Configuration ;
5 import org.apache.hadoop.fs.Path ;
6 import org.apache.hadoop.io.Text;
7 import org.apache.hadoop.mapreduce.Job ;
8 import org.apache.hadoop.mapreduce.lib.input.MultipleInputs ;
9 import org.apache.hadoop.mapreduce.lib.input.TextInputFormat ;
10 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat ;
11
12 public class DriverFile {
13     public static void main(String[] args) throws IOException, ClassNotFoundException, InterruptedException {
14         /*creating the "conf" object from the "Configuration" class, which provides access to
15         configuration parameters necessary for Hadoop job*/
16         Configuration conf = new Configuration(); // creates a new instance of the Job class named "job"
17         /*The Job.getInstance() method returns a new Job instance with the
18         specified Configuration object and a user-defined job name "Reduce-side join"*/
19         Job job=Job.getInstance() ;
20         job=Job.getInstance(conf, "Reduce-side join");
21
22
23         job.setJarByClass(DriverFile.class); //specifying the driver class in the Jar file
24         job.setReducerClass(ReducerFile.class); //specifying the reducer class in the Jar file
25         job.setOutputKeyClass(Text.class); //Setting the data type of the key output for the MapReduce job to Text
26         job.setOutputValueClass(Text.class); //setting the data type of the value output for the MapReduce job to Text
27
28
29         //Setting multiple input paths for the Hadoop MapReduce job using the MultipleInputs class.
30         MultipleInputs.addInputPath(job, new
31
32         //adding the input path of the customer data file specified by the second argument ie Customer
33         //Input format: TextInputFormat class and the Mapper function: CustomerMapperfile class
34         Path(args[1]), TextInputFormat.class, CustomerMapperfile.class);
35
36
37         // adding the input path of the transaction data file specified by the third argument ie Customer ID
38         //Input format: TextInputFormat class and the Mapper function: TransactionMapperfile class
39         MultipleInputs.addInputPath(job, new
40         Path(args[2]), TextInputFormat.class, TransactionMapperfile.class);
41         Path outputPath = new Path(args[3]); /*output path is where the final output of
42         the MapReduce job will be stored*/
43
44         //Setting the output path for Mapreduce Jobs results
45         FileOutputFormat.setOutputPath(job, outputPath);
46         // Delete the outputPath directory from the file system
47         outputPath.getFileSystem(conf).delete(outputPath);
48         //check the status of the job (completed or not)
49         System.exit(job.waitForCompletion(true)? 0:1);
50     }
51 }
```

Output key: Text  
Output value: Text

2 Mapper files  
hence Multiple  
Inputs

```
(base) hadoop@hadoop-VirtualBox:~$ hdfs namenode -format
2023-04-03 16:29:39,878 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = hadoop-VirtualBox/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.0.3
STARTUP_MSG:   classpath = /usr/local/hadoop/etc/hadoop:/usr/
/common/lib/jackson-annotations-2.7.8.jar:/usr/local/hadoop/s
/jsr311-api-1.1.1.jar:/usr/local/hadoop/share/hadoop/common/l
jar:/usr/local/hadoop/share/hadoop/common/lib/commons-lang3-3
loop/share/hadoop/common/lib/jaxb-impl-2.2.3-1.jar:/usr/local
o/share/hadoop/common/lib/jersey-server-1.19.jar:/usr/local/h
```



```
hadoop@hadoop-VirtualBox: ~  
(base) hadoop@hadoop-VirtualBox:~$ start-dfs.sh  
Starting namenodes on [localhost]  
Starting datanodes  
Starting secondary namenodes [hadoop-VirtualBox]  
(base) hadoop@hadoop-VirtualBox:~$ start-yarn.sh  
Starting resourcemanager  
Starting nodemanagers  
(base) hadoop@hadoop-VirtualBox:~$ jps  
4374 DataNode  
4619 SecondaryNameNode  
4940 ResourceManager  
5229 Jps  
4239 NameNode  
5087 NodeManager
```

Namenode information

localhost:9870/dfshealth.html#

Startup Progress

Utilities

## Overview 'localhost:9000' (active)

Started:	Mon Apr 03 16:35:17 +0100 2023
Version:	3.0.3, r37fd7d752db73d984dc31e0cddf590d252f5e075
Compiled:	Thu May 31 18:12:00 +0100 2018 by yzhang from a303
Cluster ID:	CID-c2fe915d-a941-4f8b-a362-b4fd00e7f3d1
Block Pool ID:	BP-1505303269-127.0.1.1-1680535782168

## Summary

```
hadoop@hadoop-VirtualBox: ~  
File Output Format Counters  
Bytes Written=9516  
(base) hadoop@hadoop-VirtualBox:~$ hadoop jar Downloads/HPC.jar DriverFile /input_file/CustomerData.txt /input_file/TransactionData.txt output  
_hpc  
2023-04-08 12:01:15,516 INFO impl.MetricsConfig: loaded properties from hadoop-metrics2.properties  
2023-04-08 12:01:15,874 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).  
2023-04-08 12:01:15,874 INFO impl.MetricsSystemImpl: JobTracker metrics system started  
2023-04-08 12:01:16,126 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and  
execute your application with ToolRunner to remedy this.  
2023-04-08 12:01:16,335 INFO input.FileInputFormat: Total input files to process : 1  
2023-04-08 12:01:16,516 INFO input.FileInputFormat: Total input files to process : 1  
2023-04-08 12:01:16,587 INFO mapreduce.JobSubmitter: number of splits:2  
2023-04-08 12:01:17,031 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1833845894_0001  
2023-04-08 12:01:17,034 INFO mapreduce.JobSubmitter: Executing with tokens: []  
2023-04-08 12:01:17,313 INFO mapreduce.Job: The url to track the job: http://localhost:8080/  
2023-04-08 12:01:17,315 INFO mapreduce.Job: Running job: job_local1833845894_0001  
2023-04-08 12:01:17,328 INFO mapred.LocalJobRunner: OutputCommitter set in config null  
2023-04-08 12:01:17,368 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2  
2023-04-08 12:01:17,368 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ign  
ore cleanup failures: false  
2023-04-08 12:01:17,401 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter  
2023-04-08 12:01:17,543 INFO mapred.LocalJobRunner: Waiting for map tasks  
2023-04-08 12:01:17,545 INFO mapred.LocalJobRunner: Starting task: attempt_local1833845894_0001_m_000000_0  
2023-04-08 12:01:17,644 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2  
2023-04-08 12:01:17,645 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ign  
ore cleanup failures: false  
2023-04-08 12:01:17,775 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]  
2023-04-08 12:01:17,790 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/input_file/CustomerData.txt:0+735280  
2023-04-08 12:01:17,938 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)  
2023-04-08 12:01:17,938 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100  
2023-04-08 12:01:17,938 INFO mapred.MapTask: soft limit at 83886080  
2023-04-08 12:01:17,938 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
```

### Section 3:

#### Results and Evaluations:

Customer  
Name

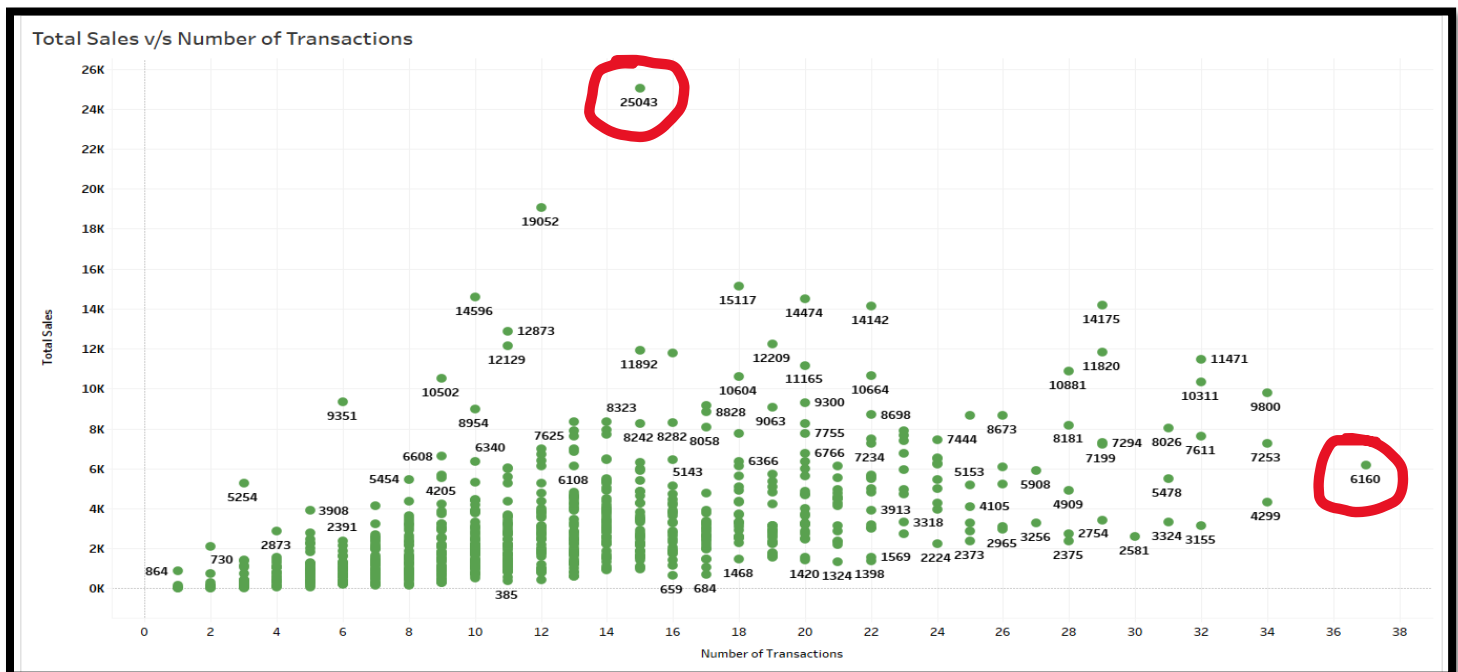
Total number of  
Transactions made  
by each Customer

Total Sales  
by each  
Customer

240	David Wiener	9	1080.750987
241	Dianna Wilson	11	5271.627948
242	Don Weiss	9	344.079999
243	Dorothy Wardle	10	1204.854001
244	Dorris liebe	6	755.601973
245	Dean percer	29	7198.760899
246	Erin Ashbrook	25	2846.705002
247	Ed Braxton	12	3644.978075
248	Edward Becker	20	3236.313920
249	Ellis Ballard	8	4358.127940
250	Emily Burns	11	2767.218974
251	Eric Barreto	7	944.599998
252	Erica Bern	6	1643.255039
253	Eugene Barchas	11	3071.125995
254	Evan Bailliet	7	1186.325989
255	Erin Creighton	7	461.913997
256	Emily Ducich	14	3484.920025
257	Emily Grady	9	1832.020026
258	Edward Hooks	32	10310.879850
259	Eric Hoffmann	17	684.171004
260	Erica Hackney	8	825.949999

#### *Output File*

Customer Names with their total number of transactions and total sales were the output of the MapReduce program which helped to answer the desired research question: What is the total amount spent by each customer on their purchases, combined? **Insights from the Output: Scatterplot**



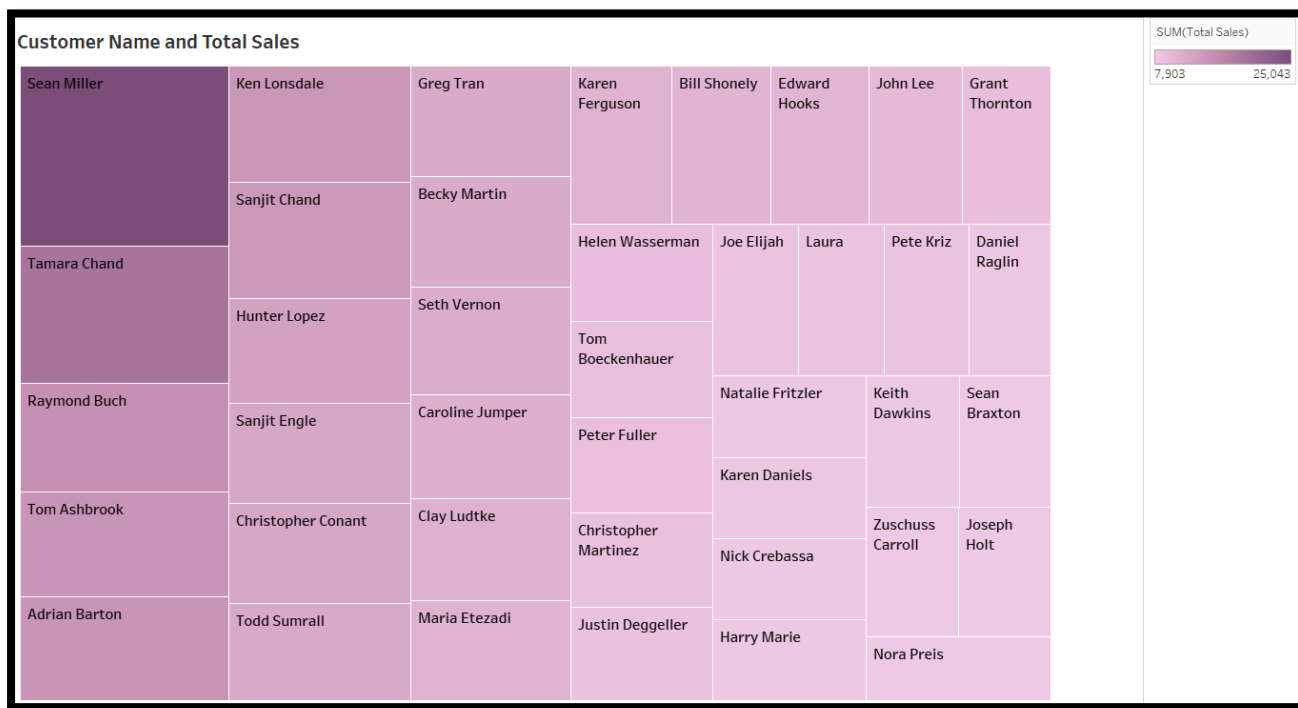
It can be seen from the above graph: **Correlation of Total Sales with the Number of Transactions** that the data points have homogenous clustering for the lower transaction numbers. The measure of the spread of the data points increases with the increase in the number of transactions. There is a positive correlation between the Total sales and the number of transactions.

However, being dependent on the number of transactions cannot help the store to predict future sales probability because it can be seen from the graph that one customer has 38(highest) transactions history but contributes to only \$6160 sales amount. On the other hand, a customer with 16 transaction history has sales values of \$25043. Finding a better variable for predicting total sales is important.

### Tree Map:

The top 40 customers contributing to maximum sales for the superstore were sampled. They were plotted on the Tree Map to understand further details and target them for customer loyalty rewards, promotions, better services, marketing, etc. The customer named Sean Miller contributed to the highest sales whereas Nora Pries contributed to one-third of the amount spent by Sean.

Further, the region of people with the highest sales can be targeted and those outlets can have more discount offers to retain and satisfy their customers. Additionally, regions, where customer sales are less, can be investigated further to formulate strategies for optimizing their customer acquisition.



### Reflection:

The rich data can be used to target the customers personally, congratulating them on their birthdays. Developing systems that communicate personalized recommendations to customers about deals and promotions. Big data can be processed faster and hence can be effectively implemented in the supermarket domain for joining two or more attributes, and for finding more hidden patterns in the dataset. Reduce side joins can be used for joining big data in the traditional HADOOP framework without the use of SQL. Additionally, the HADOOP environment can use Spark for the real-time streaming and batch processing of the superstore sales data.