Hadoop - MapReduce Queries

All java files have been compiled, thus to get the desired output simply follow the “Command to run MapReduce” section of each query documentation. Remember to ensure that the output directory does not exist before running the queries. Furthermore, note that each query uses a temporary directory to pass output of the first job into the input of the second job such as “~/temporary/Query1A”. Thus, it is advised to create a temporary directory from the home path before running the jobs. The jobs will delete their respective temporary directories before exiting. Also note to replace the value of `input/40G/store/store.dat` or `input/40G/store_sales/store_sales.dat` to the input directory of 40G store and 40G store sales database respectively.

Query 1A

Command to compile:

```
hadoop com.sun.tools.javac.Main Query1A.java
jar cf Query1A.jar Query1A*.class
```

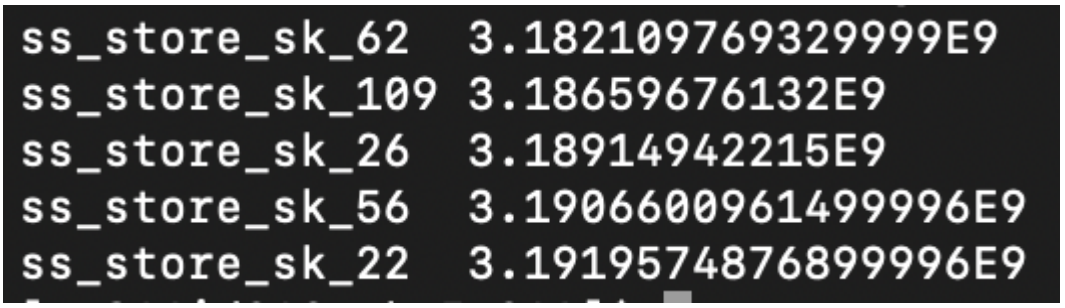
Command to run MapReduce:

```
hadoop jar Query1A.jar Query1A 5 2450816 2452642
input/40G/store_sales/store_sales.dat output/Query1A
```

Command to view output:

```
hdfs dfs -cat output/Query1A/part-r-00000
```

Sample output of running the above query:



Query 1B

Command to compile:

```
hadoop com.sun.tools.javac.Main Query1B.java
jar cf Query1B.jar Query1B*.class
```

Command to run MapReduce:

```
hadoop jar Query1B.jar Query1B 7 5 2450816 2452642
input/40G/store_sales/store_sales.dat output/Query1B
```

Command to view output:

```
hdfs dfs -cat output/Query1B/part-r-00000
```

Sample output of running the above query:

```
ss_store_sk_32  ss_item_sk_528, 128
ss_store_sk_32  ss_item_sk_2628, 116
ss_store_sk_32  ss_item_sk_24636, 113
ss_store_sk_32  ss_item_sk_23310, 104
ss_store_sk_32  ss_item_sk_36318, 99
ss_store_sk_32  ss_item_sk_23502, 91
ss_store_sk_32  ss_item_sk_13788, 74
ss_store_sk_4   ss_item_sk_43494, 174
ss_store_sk_4   ss_item_sk_13674, 149
ss_store_sk_4   ss_item_sk_37230, 148
ss_store_sk_4   ss_item_sk_38706, 144
ss_store_sk_4   ss_item_sk_15726, 131
ss_store_sk_4   ss_item_sk_48492, 122
ss_store_sk_4   ss_item_sk_22500, 113
ss_store_sk_92  ss_item_sk_46944, 133
ss_store_sk_92  ss_item_sk_9426, 127
ss_store_sk_92  ss_item_sk_39360, 120
ss_store_sk_92  ss_item_sk_46134, 115
ss_store_sk_92  ss_item_sk_51894, 101
ss_store_sk_92  ss_item_sk_6006, 97
ss_store_sk_92  ss_item_sk_49206, 95
ss_store_sk_50  ss_item_sk_44730, 142
ss_store_sk_50  ss_item_sk_26844, 137
ss_store_sk_50  ss_item_sk_3222, 128
ss_store_sk_50  ss_item_sk_4254, 122
ss_store_sk_50  ss_item_sk_26772, 120
ss_store_sk_50  ss_item_sk_50562, 101
ss_store_sk_50  ss_item_sk_19278, 86
ss_store_sk_7   ss_item_sk_40686, 149
ss_store_sk_7   ss_item_sk_46410, 135
ss_store_sk_7   ss_item_sk_16830, 117
ss_store_sk_7   ss_item_sk_32916, 115
ss_store_sk_7   ss_item_sk_32772, 110
ss_store_sk_7   ss_item_sk_16698, 99
ss_store_sk_7   ss_item_sk_36450, 94
```

Query 1C

Command to compile:

```
hadoop com.sun.tools.javac.Main Query1C.java
jar cf Query1C.jar Query1C*.class
```

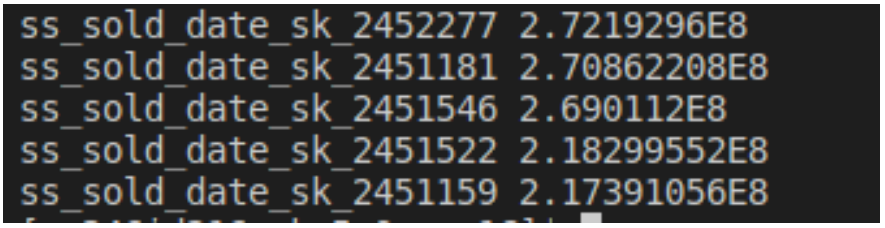
Command to run MapReduce:

```
hadoop jar Query1C.jar Query1C 5 2450816 2452642
input/40G/store_sales/store_sales.dat output/Query1C
```

Command to view output:

```
hdfs dfs -cat output/Query1C/part-r-00000
```

Sample output of running the above query:



Query 2 (Map-side Join)

Command to compile:

```
hadoop com.sun.tools.javac.Main MapJoin.java
jar cf MapJoin.jar MapJoin*.class
```

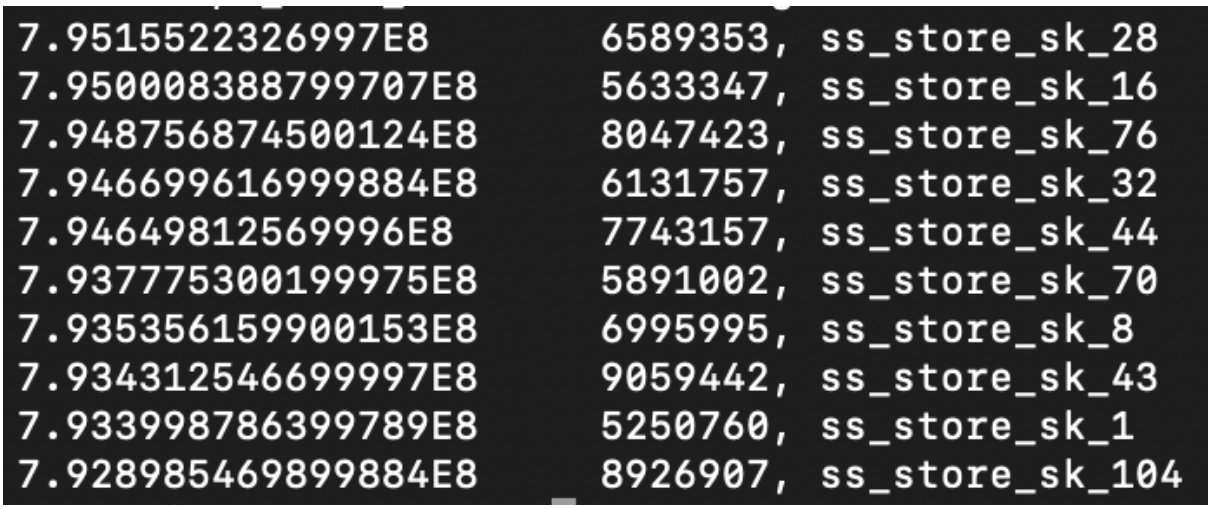
Command to run MapReduce:

```
hadoop jar MapJoin.jar MapJoin 10 2450900 2451400
input/40G/store_sales/store_sales.dat input/40G/store/store.dat
output/MapJoin
```

Command to view output:

```
hdfs dfs -cat output/MapJoin/part-r-00000
```

Sample output of running the above query:



Query 2 (Reduce-side Join)

Command to compile:

```
hadoop com.sun.tools.javac.Main ReduceJoin.java
jar cf ReduceJoin.jar ReduceJoin*.class
```

Command to run MapReduce:

```
hadoop jar ReduceJoin.jar ReduceJoin 10 2450900 2451400
input/40G/store_sales/store_sales.dat input/40G/store/store.dat
output/ReduceJoin
```

Command to view output:

```
hdfs dfs -cat output/ReduceJoin/part-r-00000
```

Sample output of running the above query:

```
7.9515522327000083E8    6589353, ss_store_sk_28
7.9500083888000101E8    5633347, ss_store_sk_16
7.9487568744998343E8    8047423, ss_store_sk_76
7.9466996170000018E8    6131757, ss_store_sk_32
7.9464981257000026E8    7743157, ss_store_sk_44
7.937775300199998E8     5891002, ss_store_sk_70
7.9353561599000072E8    6995995, ss_store_sk_8
7.9343125467000006E8    9059442, ss_store_sk_43
7.9339987863998510E8    5250760, ss_store_sk_1
7.9289854698998530E8    8926907, ss_store_sk_104
```

HiveQL Queries

External Table Schema

```
CREATE EXTERNAL TABLE store_40G(
s_store_sk int,
s_store_id string,
s_rec_start_date date,
s_rec_end_date date,
s_closed_date_sk int,
s_store_name string,
s_number_employees int,
s_floor_space int,
s_hours string,
s_manager string,
s_market_id int,
s_geography_class string,
s_market_desc string,
s_market_manager string,
s_division_id int,
s_division_name string,
s_company_id int,
s_company_name string,
s_street_number string,s_street_name string,
s_street_type string,
s_suite_number string,
s_city string,
s_county string,
s_state string,
s_zip string,
s_country string,
s_gmt_offset decimal(5,2),
s_tax_percentage decimal(5,2)
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '|'
STORED AS PARQUET
LOCATION '/user/cs346id21/input/40G/store';


CREATE EXTERNAL TABLE store_sales_40g (
ss_sold_date_sk int,
ss_sold_time_sk int,
ss_item_sk int,
ss_customer_sk int,
ss_cdemo_sk int,
```

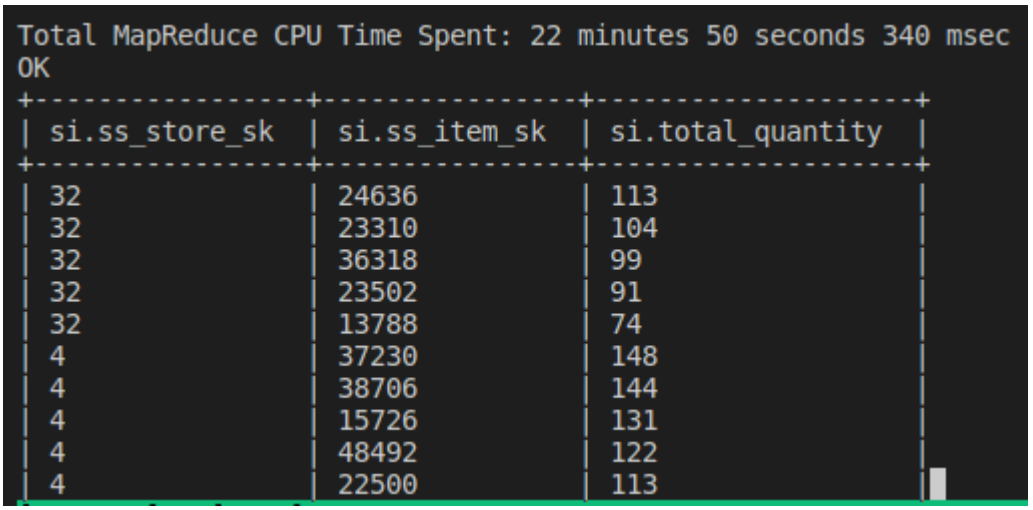
```
ss_hdemo_sk int,
ss_addr_sk int,
ss_store_sk int,
ss_promo_sk int,
ss_ticket_number int,
ss_quantity int,
ss_wholesale_cost decimal(7,2),
ss_list_price decimal(7,2),
ss_sales_price decimal(7,2),
ss_ext_discount_amt decimal(7,2),
ss_ext_sales_price decimal(7,2),
ss_ext_wholesale_cost decimal(7,2),
ss_ext_list_price decimal(7,2),
ss_ext_tax decimal(7,2),
ss_coupon_amt decimal(7,2),
ss_net_paid decimal(7,2),
ss_net_paid_inc_tax decimal(7,2),
ss_net_profit decimal(7,2)
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '|'
STORED AS PARQUET
LOCATION '/user/cs346id21/input/40G/store_sales';
```

Query 1A

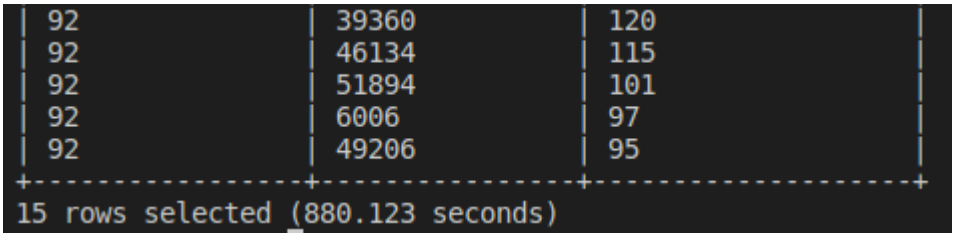
```
Total MapReduce CPU Time Spent: 9 minutes 0 seconds 810 msec
OK
+--------------+----------------+
| ss_store_sk  |    revenue     |
+--------------+----------------+
| 62           | 3182109769.33  |
| 109          | 3186596761.32  |
| 26           | 3189149422.15  |
| 56           | 3190660096.15  |
| 22           | 3191957487.69  |
+--------------+----------------+
5 rows selected (393.203 seconds)
```

```
SELECT ss_store_sk, sum(ss_net_paid) as Revenue FROM store_sales_1G
WHERE ss_store_sk IS NOT NULL and ss_net_paid IS NOT NULL and
ss_sold_date_sk IS NOT NULL and ss_sold_date_sk <= 2450816 and
ss_sold_date_sk >= 2452642 group by ss_store_sk order by Revenue ASC
LIMIT 5;
```

Query 1B



```
Total MapReduce CPU Time Spent: 22 minutes 50 seconds 340 msec
OK
+-----------------+----------------+--------------------+
| si.ss_store_sk  | si.ss_item_sk  | si.total_quantity  |
+-----------------+----------------+--------------------+
| 32              | 24636          | 113                |
| 32              | 23310          | 104                |
| 32              | 36318          | 99                 |
| 32              | 23502          | 91                 |
| 32              | 13788          | 74                 |
| 4               | 37230          | 148                |
| 4               | 38706          | 144                |
| 4               | 15726          | 131                |
| 4               | 48492          | 122                |
| 4               | 22500          | 113                |
```



```
| 92              | 39360          | 120                |
| 92              | 46134          | 115                |
| 92              | 51894          | 101                |
| 92              | 6006           | 97                 |
| 92              | 49206          | 95                 |
+-----------------+----------------+--------------------+
15 rows selected (880.123 seconds)
```

```
WITH sales_filtered AS (
  SELECT ss_store_sk, ss_item_sk, ss_quantity, ss_sold_date_sk FROM
store_sales_40G WHERE ss_sold_date_sk >= 2450816 AND ss_sold_date_sk <=
2452642
),
top_stores AS (
  SELECT ss_store_sk, SUM(ss_quantity) AS total_sold FROM sales_filtered
  GROUP BY ss_store_sk
  ORDER BY total_sold DESC
  LIMIT 3
),
top_items AS (
  SELECT ss_store_sk, ss_item_sk, SUM(ss_quantity) AS total_quantity
FROM sales_filtered WHERE ss_store_sk IN (SELECT ss_store_sk FROM
top_stores)
  GROUP BY ss_store_sk, ss_item_sk
),
store_items AS (
  SELECT ti.ss_store_sk, ti.ss_item_sk, ti.total_quantity, ts.total_sold
FROM top_items ti
  JOIN top_stores ts ON ti.ss_store_sk = ts.ss_store_sk
)
SELECT si.ss_store_sk, si.ss_item_sk, si.total_quantity
FROM (
  SELECT ss_store_sk, ss_item_sk, total_quantity, ROW_NUMBER() OVER
```

```
(PARTITION BY ss_store_sk ORDER BY total_quantity ASC) AS item_rank,
RANK() OVER (ORDER BY total_sold DESC) AS store_rank
  FROM store_items
) si
WHERE si.item_rank <= 5
ORDER BY si.store_rank ASC, si.total_quantity DESC;
```

Query 1C

```
Total MapReduce CPU Time Spent: 9 minutes 36 seconds 480 msec
OK
+------------------+---------------+
| ss_sold_date_sk  |  total_sales  |
+------------------+---------------+
| 2452277          | 272192985.36  |
| 2451181          | 270862231.93  |
| 2451546          | 269011207.69  |
| 2451522          | 218299535.32  |
| 2451159          | 217391100.15  |
+------------------+---------------+
5 rows selected (383.963 seconds)
```

```
SELECT ss_sold_date_sk, SUM(ss_net_paid_inc_tax) AS total_sales FROM
store_sales_1G WHERE ss_sold_date_sk >= 2450816 AND ss_sold_date_sk <=
2452642 GROUP BY ss_sold_date_sk ORDER BY total_sales ASC LIMIT 5;
```

Query 2

```
Total MapReduce CPU Time Spent: 11 minutes 4 seconds 350 msec
OK
+---------------+------------------+-----------------+
| s.s_store_sk  | s.s_floor_space  | total_net_paid  |
+---------------+------------------+-----------------+
| 32            | 6131757          | 3226933911.31   |
| 92            | 8573853          | 3226481169.88   |
| 50            | 7825489          | 3223393872.18   |
| 8             | 6995995          | 3219565087.71   |
| 4             | 9341467          | 3219464747.31   |
+---------------+------------------+-----------------+
5 rows selected (450.834 seconds)
```

```
SELECT s.s_store_sk, s.s_floor_space, SUM(ss.ss_net_paid) AS
total_net_paid FROM store_sales_1G ss JOIN store_1G s ON ss.ss_store_sk
= s.s_store_sk WHERE ss.ss_sold_date_sk >= 2450816 AND
ss.ss_sold_date_sk <= 2452642 GROUP BY s.s_store_sk, s.s_floor_space
ORDER BY total_net_paid DESC, s.s_floor_space DESC LIMIT 5;
```