# CS346 Coursework

Software Documentation - Group 29

## Query Commands and Outputs

Please **cd** into the respective query directory (e.g. Query 1a) before executing the Hadoop run command. Also, please ensure that each Hadoop run command is placed on a **single line** in the terminal (copying directly from this document may cause the command to be interpreted as 2 separate lines, causing an error). The Hadoop run command for each query can also be found in a comment at the top of each respective Java file (below any import statements).

### Query 1a

Hadoop run command

```
$HADOOP_HOME/bin/hadoop jar TopKNetProfit.jar TopKNetProfitDriver 10
2450816 2452642 input/40G/store_sales/store_sales.dat output/topknetprofit
```

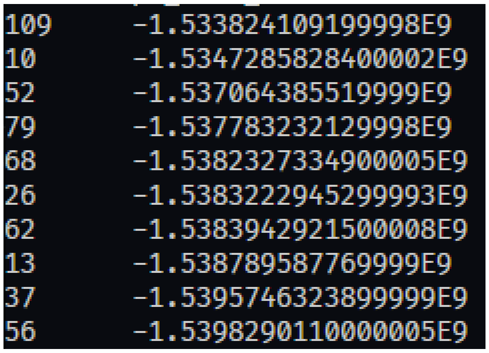
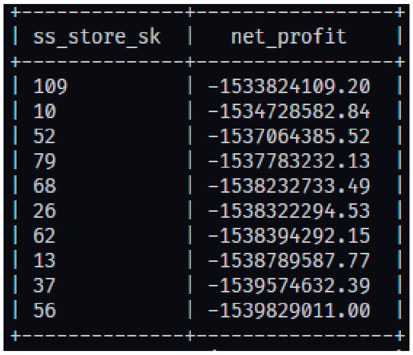
Hadoop output command

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

HiveQL Query

```
SELECT ss_store_sk,
    SUM(ss_net_profit) as net_profit
FROM store_sales_40g
WHERE ss_sold_date_sk >= 2450816
    AND ss_sold_date_sk <= 2452642
    AND ss_store_sk IS NOT NULL
    AND ss_sold_date_sk IS NOT NULL
    AND ss_net_profit IS NOT NULL
GROUP BY ss_store_sk
ORDER BY net_profit DESC
LIMIT 10;
```

Output

<table>
<tr><th>Hadoop</th><th>Hive</th></tr>
<tr><td>
<pre>
109     -1.533824109199998E9
10      -1.5347285828400002E9
52      -1.537064385519999E9
79      -1.537783232129998E9
68      -1.5382327334900005E9
26      -1.5383222945299993E9
62      -1.5383942921500008E9
13      -1.538789587769999E9
37      -1.5395746323899999E9
56      -1.5398290110000005E9
</pre></td><td>
<pre>
+--------------+-----------------+
| ss_store_sk  |   net_profit    |
+--------------+-----------------+
| 109          | -1533824109.20  |
| 10           | -1534728582.84  |
| 52           | -1537064385.52  |
| 79           | -1537783232.13  |
| 68           | -1538232733.49  |
| 26           | -1538322294.53  |
| 62           | -1538394292.15  |
| 13           | -1538789587.77  |
| 37           | -1539574632.39  |
| 56           | -1539829011.00  |
+--------------+-----------------+
</pre></td></tr>
</table>

## Query 1b

Hadoop run command

```
$HADOOP_HOME/bin/hadoop jar TopKSoldItems.jar TopKSoldItemsDriver 10
2450816 2452642 input/40G/store_sales/store_sales.dat output/topksolditems
```

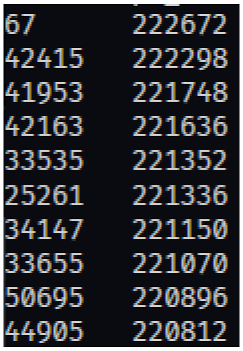
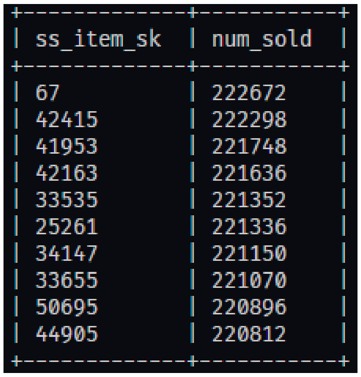
Hadoop output command

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

HiveQL Query

```
SELECT ss_item_sk,
     SUM(ss_quantity) as num_sold
FROM store_sales_40g
WHERE ss_sold_date_sk >= 2450816
     AND ss_sold_date_sk <= 2452642
     AND ss_item_sk IS NOT NULL
GROUP BY ss_item_sk
ORDER BY num_sold DESC
LIMIT 10;
```

Output

| Hadoop | Hive |
|---|---|
| <br>67 222672<br>42415 222298<br>41953 221748<br>42163 221636<br>33535 221352<br>25261 221336<br>34147 221150<br>33655 221070<br>50695 220896<br>44905 220812 | <br>+-------------+-----------+<br>\| ss_item_sk \| num_sold \|<br>+-------------+-----------+<br>\| 67 \| 222672 \|<br>\| 42415 \| 222298 \|<br>\| 41953 \| 221748 \|<br>\| 42163 \| 221636 \|<br>\| 33535 \| 221352 \|<br>\| 25261 \| 221336 \|<br>\| 34147 \| 221150 \|<br>\| 33655 \| 221070 \|<br>\| 50695 \| 220896 \|<br>\| 44905 \| 220812 \|<br>+-------------+-----------+ |

**Query 1c**

Hadoop run command

```
$HADOOP_HOME/bin/hadoop jar TopKNetProfitByDate.jar
TopKNetProfitByDateDriver 10 2451520 2451771
input/40G/store_sales/store_sales.dat output/topknetprofitdate
```

Hadoop output command

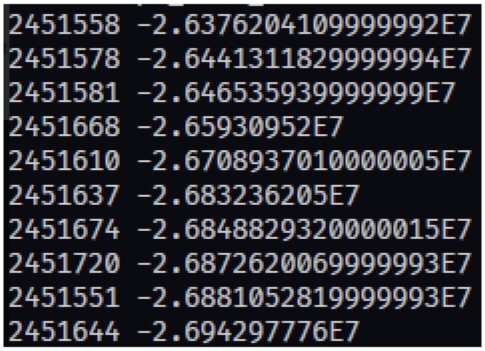
```
hdfs dfs -cat output/topknetprofitdate/part-r-00000
```

HiveQL Query

```
SELECT ss_sold_date_sk,
    SUM(ss_net_profit) as net_profit
FROM store_sales_40g
WHERE ss_sold_date_sk >= 2451520
    AND ss_sold_date_sk <= 2451771
    AND ss_sold_date_sk IS NOT NULL
    AND ss_net_profit IS NOT NULL
GROUP BY ss_sold_date_sk
ORDER BY net_profit DESC
LIMIT 10;
```

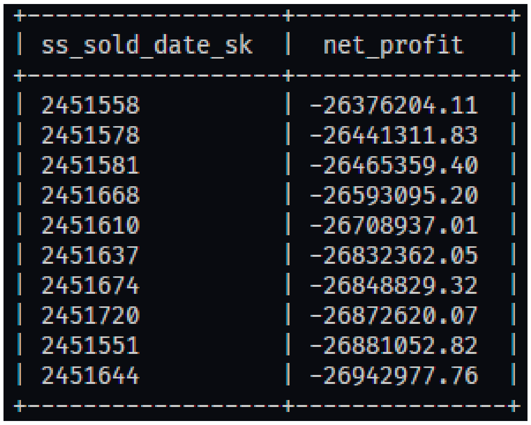
Output

<table>
<tr><th>Hadoop</th><th>Hive</th></tr>
<tr><td>



```
2451558 -2.6376204109999992E7
2451578 -2.6441311829999994E7
2451581 -2.646535939999999E7
2451668 -2.65930952E7
2451610 -2.6708937010000005E7
2451637 -2.683236205E7
2451674 -2.6848829320000015E7
2451720 -2.6872620069999993E7
2451551 -2.6881052819999993E7
2451644 -2.694297776E7
```

</td><td>



```
+-----------------+---------------+
| ss_sold_date_sk |  net_profit   |
+-----------------+---------------+
| 2451558         | -26376204.11  |
| 2451578         | -26441311.83  |
| 2451581         | -26465359.40  |
| 2451668         | -26593095.20  |
| 2451610         | -26708937.01  |
| 2451637         | -26832362.05  |
| 2451674         | -26848829.32  |
| 2451720         | -26872620.07  |
| 2451551         | -26881052.82  |
| 2451644         | -26942977.76  |
+-----------------+---------------+
```

</td></tr>
</table>

## Query 2

Hadoop run command

```
$HADOOP_HOME/bin/hadoop jar TopKStoreProfitEmployees.jar
TopKStoreProfitEmployeesDriver 10 2450816 2452642
input/40G/store_sales/store_sales.dat input/1G/store/store.dat
output/join_result
```

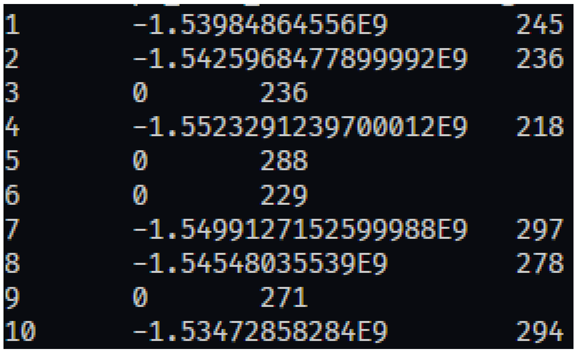
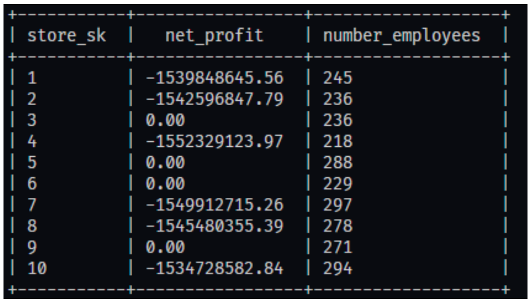
Hadoop output command

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

HiveQL Query

```
SELECT b.s_store_sk AS store_sk,
    COALESCE(a.net_profit, 0) AS net_profit,
    b.s_number_employees AS number_employees
FROM (
    SELECT ss_store_sk,
        SUM(ss_net_profit) as net_profit
    FROM store_sales_40g
    WHERE ss_sold_date_sk >= 2450816
        AND ss_sold_date_sk <= 2452642
        AND ss_store_sk IS NOT NULL
        AND ss_sold_date_sk IS NOT NULL
        AND ss_net_profit IS NOT NULL
    GROUP BY ss_store_sk
) a RIGHT OUTER JOIN (
    SELECT s_store_sk,
        s_number_employees
    FROM store_40g
    WHERE s_number_employees IS NOT NULL
) b ON a.ss_store_sk = b.s_store_sk
ORDER BY b.s_store_sk ASC
LIMIT 10;
```

Output

| Hadoop | Hive |
|---|---|
|  |  |

```
1       -1.53984864556E9        245
2       -1.5425968477899992E9   236
3       0       236
4       -1.5523291239700012E9   218
5       0       288
6       0       229
7       -1.5499127152599988E9   297
8       -1.54548035539E9        278
9       0       271
10      -1.53472858284E9        294
```

```
+-----------+-----------------+--------------------+
| store_sk  |   net_profit    | number_employees   |
+-----------+-----------------+--------------------+
| 1         | -1539848645.56  | 245                |
| 2         | -1542596847.79  | 236                |
| 3         | 0.00            | 236                |
| 4         | -1552329123.97  | 218                |
| 5         | 0.00            | 288                |
| 6         | 0.00            | 229                |
| 7         | -1549912715.26  | 297                |
| 8         | -1545480355.39  | 278                |
| 9         | 0.00            | 271                |
| 10        | -1534728582.84  | 294                |
+-----------+-----------------+--------------------+
```

## Hive External Table Creation Commands

### store_sales.dat

```
create external table if not exists store_sales_40G(
    ss_sold_date_sk bigint,
    ss_sold_time_sk bigint,
    ss_item_sk bigint,
    ss_customer_sk bigint,
    ss_cdemo_sk bigint,
    ss_hdemo_sk bigint,
    ss_addr_sk bigint,
    ss_store_sk bigint,
    ss_promo_sk bigint,
    ss_ticket_number bigint,
    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 '|'
location 'input/40G/store_sales/';
```

stores.dat

```
create external table if not exists store_40G(
      s_store_sk bigint,
      s_store_id char(16),
      s_rec_start_date date,
      s_rec_end_date date,
      s_closed_date_sk bigint,
      s_store_name varchar(50),
      s_number_employees int,
      s_floor_space int,
      s_hours char(20),
      S_manager varchar(40),
      S_market_id int,
      S_geography_class varchar(100),
      S_market_desc varchar(100),
      s_market_manager varchar(40),
      s_division_id int,
      s_division_name varchar(50),
      s_company_id int,
      s_company_name varchar(50),
      s_street_number varchar(10),
      s_street_name varchar(60),
      s_street_type char(15),
      s_suite_number char(10),
      s_city varchar(60),
      s_county varchar(30),
      s_state char(2),
      s_zip char(10),
      s_country varchar(20),
      s_gmt_offset decimal(5,2),
      s_tax_percentage decimal(5,2)
)
row format delimited fields terminated by '|'
location 'input/40G/store/';
```

## Design Decisions

An in-depth discussion on the design and implementation decisions made can be found within the coursework report. This includes the topics:

- Hadoop design patterns and implementation
- HiveQL query design and implementation
- HiveQL table creation description
- Optimisation approaches and results