# Data Engineer Practical Test Q2

Some insights gained from the analysis:

1. For Data cleaning I have dropped the column Customer Id as it had a lot of missing values and wasn't needed for analysis. I have also added the sales amount column in the dataset
2. I have also filled the null values in description column with Unknown item.
3. The results of the questions asked are below:

   **Total Sales and Number of Transactions per day:**

```
: sales_summary.show()
```

```
[Stage 41:==============>
+---------+-----------+----------------------+
|  InvDate|total_sales|number_of_transactions|
+---------+-----------+----------------------+
|3/23/2011|      16558|                  1319|
|2/16/2011|      17046|                  1191|
|2/10/2011|       8827|                   785|
|3/13/2011|       2664|                   537|
|1/20/2011|      12447|                  1502|
| 2/9/2011|      13033|                   879|
|3/25/2011|      23030|                  1386|
|1/30/2011|       4465|                   722|
| 2/8/2011|      15597|                  1228|
|1/17/2011|      21989|                  2557|
|1/10/2011|      17498|                  1976|
|1/13/2011|      14732|                  1445|
|3/21/2011|      10649|                  1068|
|1/16/2011|       4887|                   646|
|1/18/2011|      14378|                  1447|
|2/21/2011|      12567|                  1425|
| 1/5/2011|     -11751|                  1743|
|2/27/2011|       6564|                   812|
|1/12/2011|      18326|                  1809|
|2/13/2011|       4068|                   624|
+---------+-----------+----------------------+
only showing top 20 rows
```

**Top 10 products with the Highest sales:**

| Description | total_sales |
|---|---|
| DOTCOM POSTAGE | 205896 |
| REGENCY CAKESTAND... | 153384 |
| PARTY BUNTING | 87526 |
| WHITE HANGING HEA... | 77333 |
| POSTAGE | 66254 |
| JUMBO BAG RED RET... | 65022 |
| PAPER CHAIN KIT 5... | 50053 |
| RABBIT NIGHT LIGHT | 48719 |
| CHILLI LIGHTS | 46149 |
| PICNIC BASKET WIC... | 39589 |