**Identification of Duplicate Records**

This is the data processing task where we identify duplicate records across the entire data set. In some cases, we load the data what we get from our vendor / parter to a table or a file. Consider the customer\_input is the table / file where we got data.

You have to populate the customer\_target table / data set with the unique records. If there are any duplicate records then you should store the same in the other table with the number of duplicate records.

|  |  |  |  |
| --- | --- | --- | --- |
| **Customer\_input** |  |  |  |
| **cust\_name** | **cust\_phone** | **cust\_city** | **cust\_since** |
| TIM | 12345678 | Chennai | 15-Jan-05 |
| BALA | 49595089 | Chennai | 16-Jan-06 |
| RAMESH | 345093450 | Mumbai | 17-Feb-07 |
| BALA | 49595089 | Chennai | 16-Jan-06 |
| TIM | 12345678 | Chennai | 15-Jan-05 |
| BALA | 49595089 | Chennai | 16-Jan-06 |

**Output table**

|  |  |  |  |
| --- | --- | --- | --- |
| **cust\_target** |  |  |  |
| **cust\_name** | **cust\_phone** | **cust\_city** | **cust\_since** |
| TIM | 12345678 | Chennai | 15-Jan-05 |
| BALA | 49595089 | Chennai | 16-Jan-06 |
| RAMESH | 345093450 | Mumbai | 17-Feb-07 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **cust\_duplicate** |  |  |  |  |
| **cust\_name** | **cust\_phone** | **cust\_city** | **cust\_since** | **no\_of\_instances** |
| TIM | 12345678 | Chennai | 15-Jan-05 | 2 |
| BALA | 49595089 | Chennai | 16-Jan-06 | 3 |

Create data pipeline / procedure / job to accomplish this activity.