This is the data processing task where we identify duplicate records across the entire file. In some of the files we get from the other business partner, typically we don’t get any unique key. In this cases, we have to identify the duplicates so that your data processing goes with out any issues to populate the DW.

Consider the first file as input. We can have this any directory, so your program should consider taking the input file as the argument.

You have to create two files as output.

First file (you should be able to send this argument) should have the unique records and the second file (you should be able to send this as argument) should have the duplicate records and the no of records.

Create the ETL job based on your tool selection.

|  |  |  |  |
| --- | --- | --- | --- |
| **customer\_file** |  |  |  |
| **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 |
| TIM | 12345678 | Chennai | 15-Jan-05 |
| BALA | 49595089 | Chennai | 16-Jan-06 |

Create table customer

(cust\_name varchar2(20),

Cust\_phonenumber(10),

Cust\_city varchar2(20),

Cust\_since date

)

**Output files what I need**

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

Create table cust\_unique

(cust\_name varchar2(20),

Cust\_phone number(10),

Cust\_city varchar2(20),

Cust\_since date);

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 | 2 |

Create table cust\_duplicate

(cust\_name varchar2(20),

Cust\_phone number(10),

Cust\_city varchar2(20),

Cust\_since date,

No\_of\_instances number(4)

)

Tip: If you are using Stored procedure to solve then you can consider tables instead of files. If you are able to use the file names, that is great.

ETL tool: Some where in the tool or in configuration table I should be able to mention the location and file names which are involved in this job.

insert into customer values('raj',12345,'blr','21-jan-19');

insert into customer values('kiran',989734,'hyd','19-mar-18');

insert into customer values('kimm',878384,'chn','29-apr-19');

insert into customer values('rajesh',783738,'blr','31-jan-18');

insert into customer values('arun',892393,'del','23-aug-18');

insert into customer values('rajesh',783738,'blr','31-jan-18');

insert into customer values('arun',892393,'del','23-aug-18');