# ETL Process Modeling In DWH Using Enhanced Quality Techniques

*Kushanoor Akbar[1], Dr. S.Murali Krishna[2] and T. Vidya Sagar Reddy[3]*

*[1]JNTUA, Anantapur university, Madanapalle Institute of Technology & Science, Kadiri Road Angallu (Vill), Madanapalle-517 325, India*
*[2] JNTUA, Anantapur university, Professor and Head in CSE Dept, Madanapalle Institute of Technology & Science, Kadiri Road Angallu (Vill), Madanapalle-517 325, India*
*[3]Senior Data Warehouse consultant, Capgemini PVT LTD, Bangalore, India*
*[1]akbar.akbar8@gmail.com, [2]drmuralimits@gmail.com, [3]vidyaveda4@gmail.com*

***Abstract***

*Large organizations have a lot of data. The data can be stored in many formats including data bases and unstructured file. This data bases must be collected, compared and made to work as a seamless whole but the different databases communicate well. A Data warehouse is an integrated collection of subject- oriented data in the support of decision making. The integration of data sources is achieved through the use of ETL (Extract, Transform and load) process. It is therefore extensively recognized that the appropriate design of ETL process are key factors in success of Data Warehouse Project. Data warehouse is used to provide effective result from multi- dimensional data analysis. Defective data lead to break downs in the supply chain, poor business decisions and inferior customer relationship management. So data quality is the degree to which data meet the specific needs of the customer. The accuracy and correctness of the results depend on the quality of the data. Improving the quality of data is important in data warehouse because it is used in the process decision support which requires accurate data. This project presents a data warehouse construction with quality decision support system to "Manage results for an organization using customer care center". Organization used to maintain customer care to support and handle customer queries, to maintain details of customers, to provide frequent information regarding to their premiums, loans. This project determines a detail report such as how many customers are there in an Organization. How many customers paid full premiums, what are their dues, total amount paid? Which locations customer exists? How many customers are more valued customers? Total amount credited in organizations quarterly, what percent is gain/loss. In this paper we take source as flat files, relational tables and the data is extracted in staging area and then it is loaded in to a data warehouse. The different five themes frame our analysis is: Integration, Implementation, Intelligence, and Innovation and quality. The factors Definition conformance, completeness, validity, accuracy, non- duplication, accessibility applied on data warehouse dynamically to improve the performance of data warehouses.*

***Keywords:*** *ETL, Decision support System, Informatica, Oracle, Flat Files, Quality Factors*

## 1. Introduction

In the 90' the term Data warehouse with Decision supports to indicate a combination of software, database, some methodologies and analytical tools. The system is designed to give right information at right time in the right place. This system aims to help organizations in the processing of large amount of heterogeneous data. The data extracted manually from different information system is very difficult for formulate strategies and tactics for effective and

profitable business. For a decade organization takes business decision, these decisions will take on basis of analysis made mainly on the structural part of the available data. Particular Data warehousing systems represent the decision support system on which industry focus on the particular problem. The information system in the organizations generally divided into two categories. 1. Operational supporting system (OSS). 2. Decision supporting system (DSS). OSS serves the need of running the daily operations of the business. DSS provide historical information for analyzing to the business. So that important decisions can be made appropriately. Data warehouse is a subject-oriented, integrated, time variant & non-volatile collection of data in support management's decision making process. ETL layer is one of the most important layers in data warehouse scheme. In ETL data is collected, integrated, transformed & load into the Data warehouse & proved vice management's to single analysis environments to help them carry out scientific decisions. We believed that 75% of information of any organizations is contained in unstructured & semi structured data. The data extraction process starts with identification of the heterogonous databases & their format of representation. The information transformed according to business need that is change format, filtering, and merging data from multiple sources & usually in databases. The aggregation of information helps to improve queries & to improve the performance of the Data Warehouse. The data extracted from sources that will be stored in a repository of consistent historical data that can be easily make as a single source.

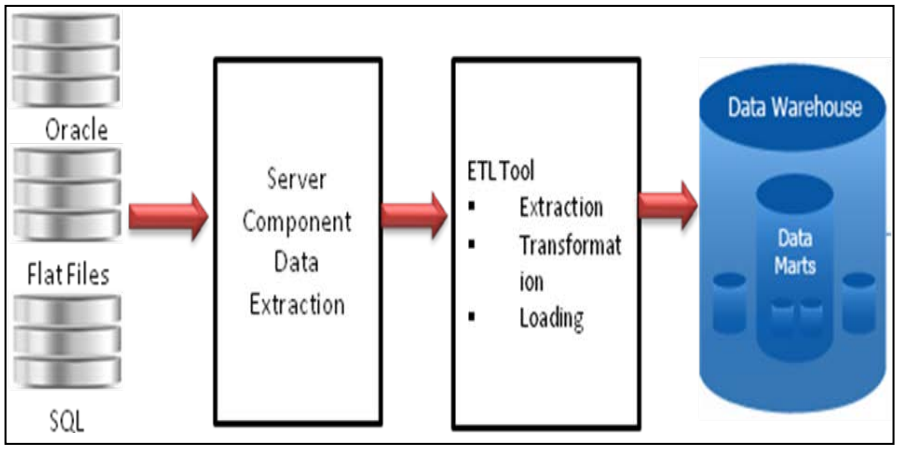
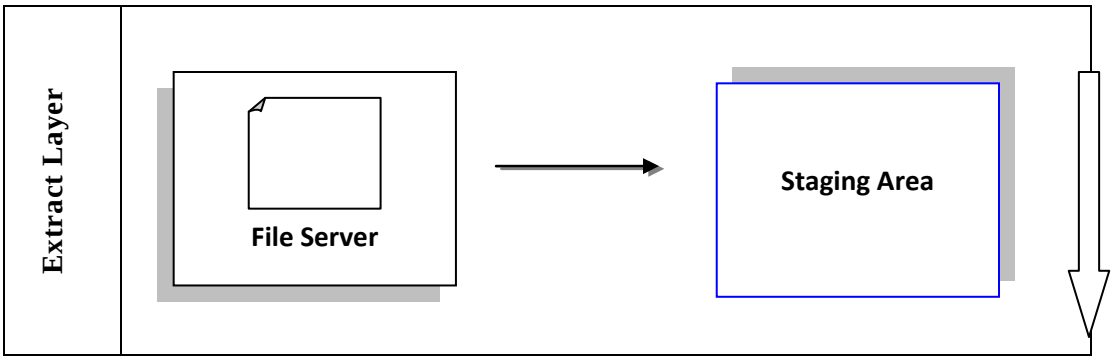



**Figure 1. Architecture of Data Warehouse**

## 2. ETL Overview



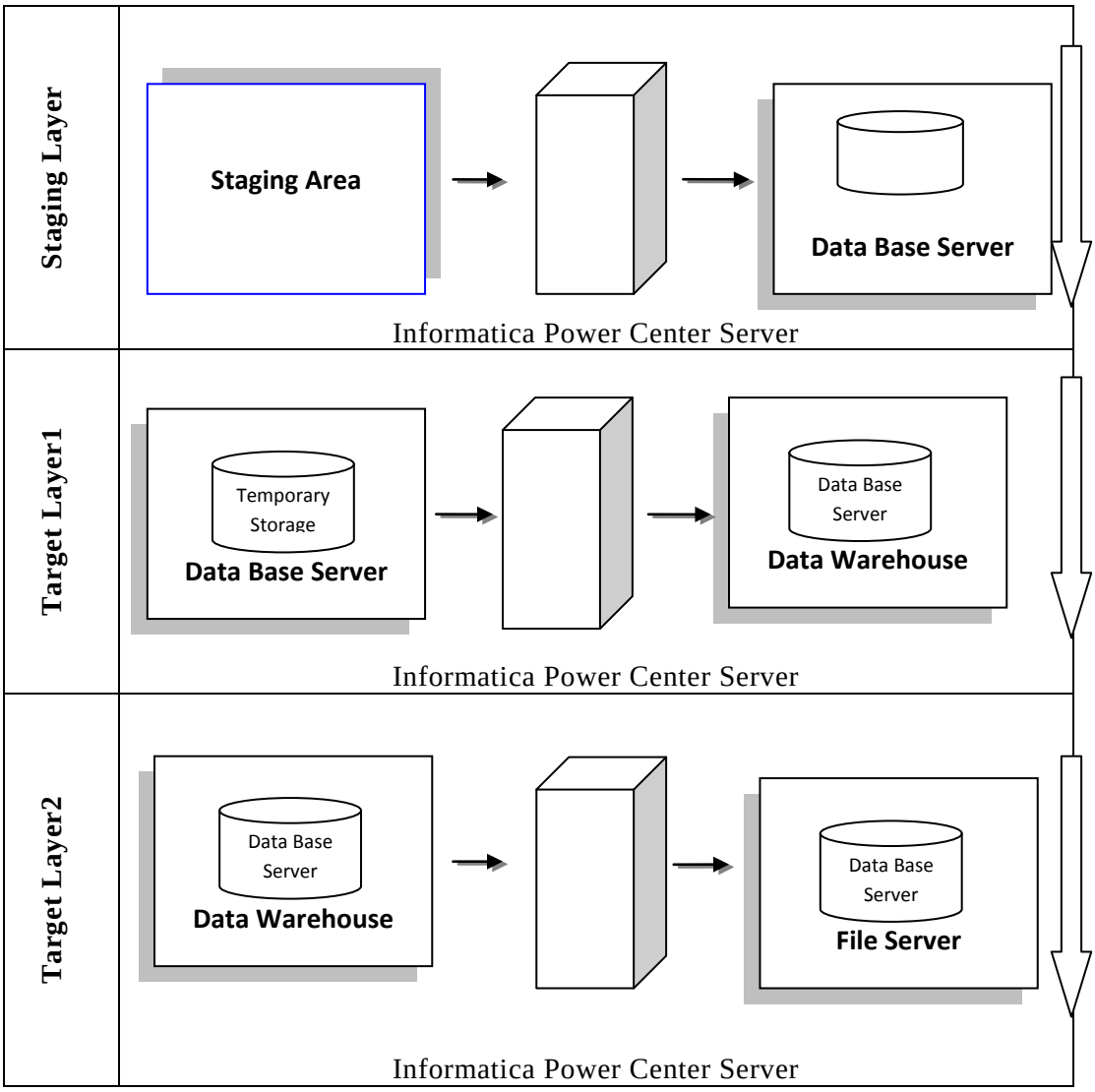



**Figure 2. ETL Process Flow for Building the DWH**

The ETL Process has two phases in high level.

1. Extracting and Loading the Files received from Transaction based automated system Team to DWH.

2. Creating and Uploading Raw Data File feeds from DWH to various Business Users.

ETL process is used to load the source files into a target oracle data warehouse. To maintain restorability and load the complete subject area in least time. ETL has been broken down into three areas based on the ETL target area. Following are the main areas:

**1.** Extract Layer.

**2.** Staging Layer.

3. Target Layer1.
4. Target Layer2.

The target layer can be divided into two phases to support the data warehouse and file archiving or file feed generation. Now we discuss the each layer.

### 2.1. Extract Layer:

The source files are extracted from the source system to ETL environment. The files are loaded without any logical transformation done on them.

### 2.2. Staging Layer:

Staging layer is an oracle database which is almost a replica of the file layout. In this layer data is loaded for temporally for daily snapshot of fact table.

### 2.3. Target Layer1:

The data loaded in the staging tables are passed on as direct load on the actual target fact table. In this layer we applying the business rules to the data. Here business rules nothing but as per the user requirement the data is modified or updated or transformed.

### 2.4. Target Layer2:

File are taking backup here the generated files will be uploaded to database. Source files are extracted using mapping from the source system. File validation process will be executed and file validated after the extraction. Once file validation is completed, source to staging sessions will be executed and data loaded into the staging area. Staging to data warehouse process will be executed and data loaded into data warehouse with some business logic. Archiving of the processed files is done using scripts or power center.

## 3. File Naming Convention

Each file in the data warehouse has naming conventions for easily identifying the table from the list of tables. Here TS represent the transaction file name with the extension of file name and data of the fillies created. Count file is used to represent the number file generated within a day, week, month, year. So for easy identification of count file the naming convention is “DaileyBalanceYYYYMMDD **HH .txt”** Target Table start with T_source File name. Mapping name convention is **M_Target_File_Name_Source_File**. Here **M** represents the mapping, **Target_file_name** represents the name of the target file, **and Source_file** represents the name of the source file.

## 4. Mapping Specification

The Mapping Specification contains the details of data move from source to Data Warehouse. The details of mapping name give the clear details of the where the data is extracted and where to load the data. Mapping Specification also contains the table for processing the data form source to Data Warehouse.

| File | Naming Convention |
|---|---|
| **Transaction Files** | TS_Filename_Date |
| **Target File** | T_File_Type_Name |
| **Mapping** | M_Target_File_Name_Source_File |
| **Count File** | Daily BalanceYYYYMMDDHH.txt |

**Figure 3. File Naming Convention**

| Mapping Name: | M_Source_Stagging_Target_Data_Dictionary |
|---|---|
| **Source System(s):** | **Success_filelist_data_dictionary_file.dat** |
| **Target System(s):** | **Data_Dictionary_STG** |
| **Initial Rows:** | As this mapping reads flat file and writes to table, if file not exist, job will not be kicked off. So there is no necessary to have initial rows. |
| **Short Description:** | To extract **Data dictionary (Generic Variables)** from source and load it into the Data_Dictionary_STG tables. |
| **Load Frequency:** | Daily/Weekly/Monthly/Quarterly/Yearly. |
| **Error Strategy:** | Invalid records will be sent to the reject table triggering a mail at the end of process. |
| **Reload Strategy:** | Restart the session from where it failed. |
| **Dependant Objects** | File Validation session. |

**Figure 4. Mapping Specification**

### 4.1. Sources

| Files | | | |
|---|---|---|---|
| **File Name** | **File Location** | **Fixed/Delimited** | **Additional File Info** |
| **success_filelist_data_dictionary_file.dat** | Source System (Oracle) | Fixed | NA |

### 4.2. Targets

| Tables | Schema Owner | | | |
|---|---|---|---|---|
| **Table Name** | **Update** | **Delete** | **Insert** | **Unique Key** |
| **Data_Dictionary_STG** | | | **Insert** | **EMP_ID** |

## 5. Auditing and Balancing

ETL extract will populate specific tables mentioned below to support automated front end balancing of daily file feeds from source to extract, staging and load. The record count for the source system by using the trailer count of file, that is number of records received in the trailer count is matched with the actual number of detailed records received using a file validation mapping. All the details are maintained into the table that is audit table.

| NAME | DATA TYPE | LENGTH | NULLABLE |
|---|---|---|---|
| **Load_id** | INTEGER | 4 | NO |
| **Source_NM** | CHARACTER | 50 | NO |
| **Session_NM** | CHARACTER | 50 | NO |
| **Ins_REC** | INTEGER | 40 | NO |
| **Rje_it** | INTEGER | 4 | NO |
| **FILE_SEQ_NO** | INTEGER | 4 | NO |

**Figure 5. Auditing Table**

## 6. Reasons for Holding the ETL Process

The ETL process will be put on hold

- File doesn't balance.
- File doesn't available in the source system.
- Files fail validation.
- All the Daily / Adhoc files are not received.
- Target file is not available.

## 7. Error Detection and Capture Design

The design for error detection and capture that will be implemented for the data warehouse projects. Error detection will need to happen at various levels since errors originate in various phases of the ETL process.

The following types of errors need to be captured.

### 7.1. Transformation Errors

Errors encountered by the power center server while transforming data. This type of errors appearing in the tool what we are using for the construction of data warehouse.

### 7.2. Business Logic Errors

These errors will be, due to violating business rules or logic errors as defined in the tool mappings based on requirements from the business. The following table maps different error types to error classification and defines the method for deleting and capturing the errors.

| Type of Error | Error Classification | Method Captured |
|---|---|---|
| Transformation Error | Technical Rejects | Error Log File |
| Business Logic Error | User Rejects (or) Business Rejects | Table Rejects |

**Figure 6. Error Types**

## 8. Dimensional Modeling in Data Warehouses

To overcome the performance issues for large queries in the data warehouse we use the dimensional models. The dimensional modeling improves the query performance for summary reports without affecting data. A dimensional data base requires much more space than its relational data base. Dimensional model consist of two types of tables having different characteristics.

1. Fact Tables.
2. Dimensional Tables.

### 8.1.1. Fact Table

Fact table contains numerical values of what we measures. Each fact table contains the key to associated dimension table. These are referred as keys; these keys are called **foreign keys.**

### 8.1.2. Dimensional Table

It contains the details about facts. The dimensional table contains descriptive information about the numerical values in the fact table. It contains the attributes of Facts. Data in a dimensional table is **De normalized**.

### 8.2. Star Schema

It has on fact table and several dimension tables. The dimension table is not denormalized. Start schema consists of fact table that are nothing but numerical values that are additive measurements and dimensional table contain smaller and descriptive data. In the figure represent the star schema of ORDER. Here ORDER FACTS contain the numerical values of customer, product, sale person, order Type, so it is called as fact table. The remaining tables contain the descriptive values of the facts so they are all called as a Dimensional Tables.
**Fact Table→** ORDER_FACTS

**Dimensional Tables →** CUSTOMER, PRODUCT, SALESPERSON, ORDER_TYPE.

### 8.3. Snowflake Schema

Further normalization and expansion of the dimension tables in a star schema is called snowflake schema. A dimension is said to be snowflake when the low-cardinality columns in the dimension have been removed to separate normalized tables that then link back into the original dimensional table. In the figure represent the snowflake schema of ORDER. Here ORDER FACTS contain the numerical values of customer, product, sale person, order Type, so it is called as fact table. The remaining tables contain the descriptive values of the facts so they are all called as a Dimensional Tables. The product table contains the brand dimension. The Brand dimension normalized into separate dimension table. Same as Customer dimensional table contain the industry dimension it also separate normalized to separate dimensional table. So it meets the snowflake schema requirement.

## 9. Data Warehouse Quality Factors

Data warehouse is used to provide effective results from multidimensional data analysis. The accuracy and correctness of these results depend on the quality of the data. Data quality is the degree to which data meet the specific needs of specific customers, which contains several dimensions. Poor data quality costs business vast amount of money every year. Defective data lead to breakdowns in the supply chain, poor business decisions and inferior customer relationship management. Data are the core business asset that needs to be managed if an organization is to generate a return from it.

The following are characteristics and measures of Data Quality:

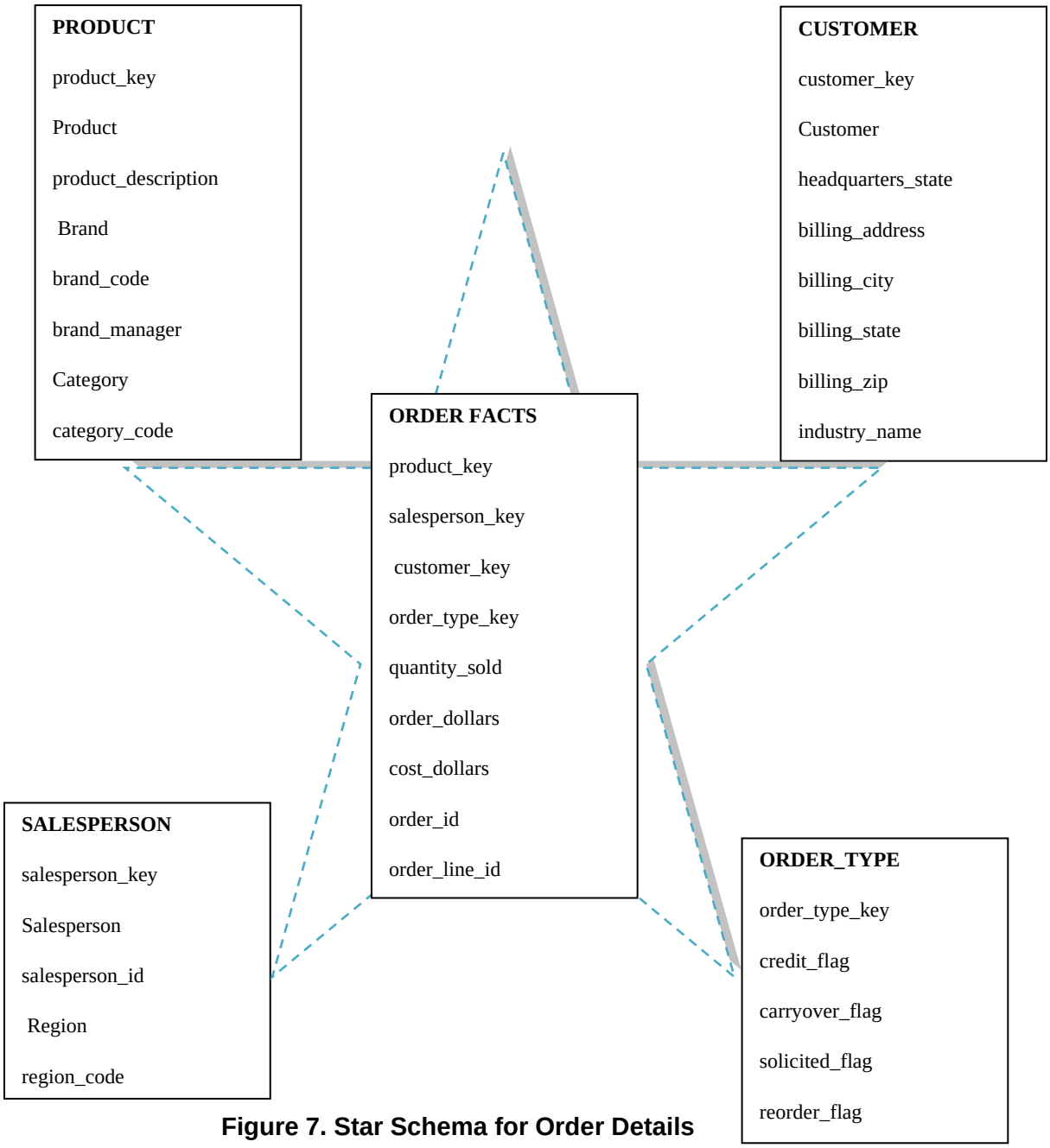



**Figure 7. Star Schema for Order Details**

1. Definition conformance
2. Completeness (Of Values)
3. Validity (Business Rule Conformance)
4. Accuracy (To the Source)
5. Non- Duplication
6. Accessibility

**1. Definition Conformance**

The chosen object is of most important and its definition should have complete details and meaning of the real world object.

**2. Completeness (Of Values)**

It is the characteristics of having all required values for the data fields.

**3. Validity (Business rule conformance)**

It is a measure of degree of conformance of data values to its domain and business rules. These include domain values, range, reasonability tests, primary key uniqueness, referential integrity.

**4. Accuracy (To the source)**

It is a measure of the degree to which data agrees with data contains in an original source.

**5. Non Duplication**

It is the degree to which that is a one- to – one correlation between records and the real world object (or) events being represented.

**6. Accessibility**

Is the characteristic of being able to access data on demand?

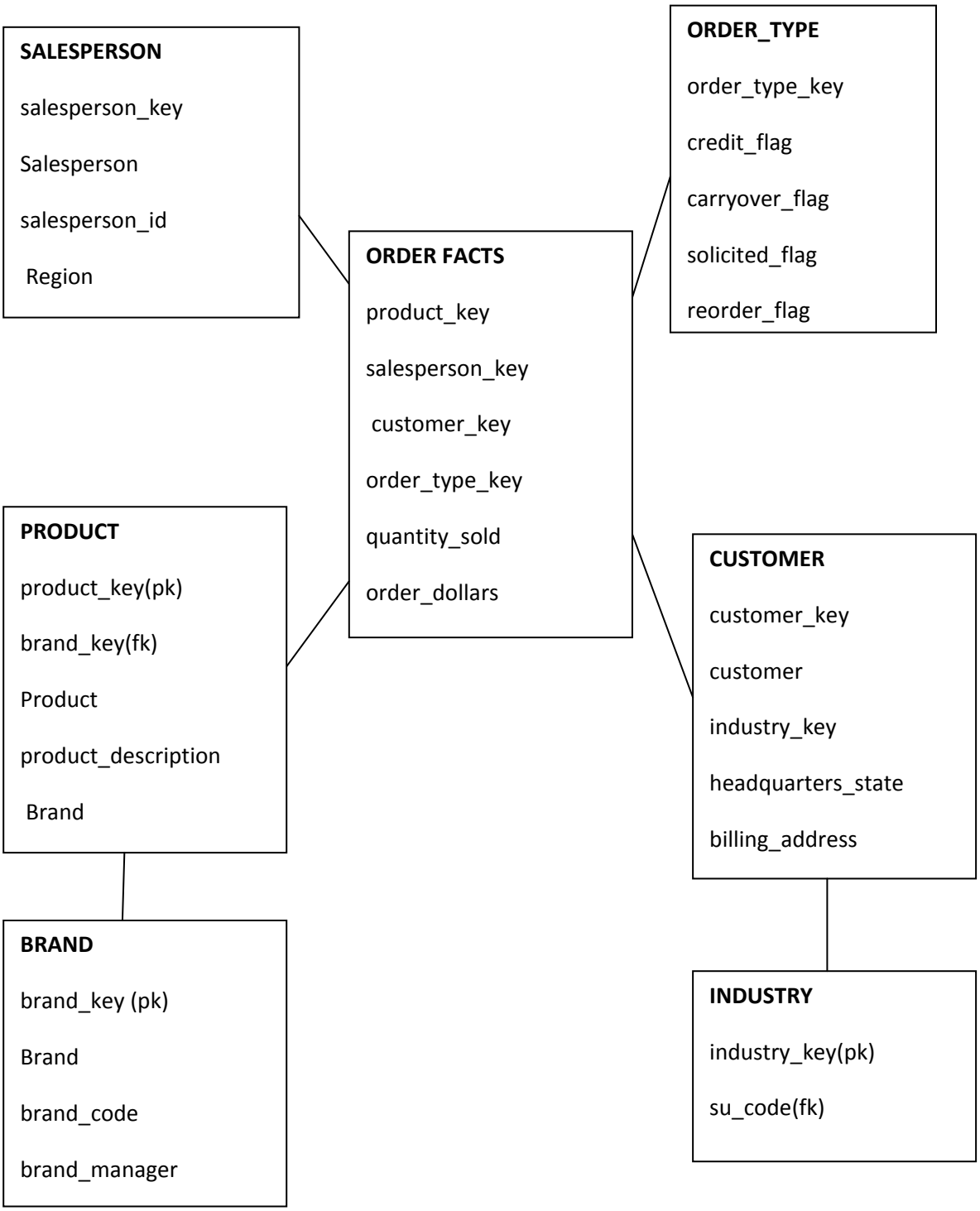



**Figure 8. Snowflake Schema for Order Details**

# 10. Results

## 10.1. Star Schema

## 10.2. Completeness (Of Values)

It is the characteristics of having all required values for the data fields.

### 10.2.1. Applies to:

Informatica Power Center 8.6.0

### 10.2.3. Summary

It describes a solution to fill the null values to gain the required values for the data fields.

**Table of Contents**



## 10.1. Star Schema

It has on fact table and several dimension tables. The dimension table is not denormalized. Start schema consists of fact table that are nothing but numerical values that are additive measurements and dimensional table contain smaller and descriptive data. In the figure represent the star schema of ORDER. Here ORDER FACTS contain the numerical values of customer, product, sale person, order Type, so it is called as fact table. The remaining tables contain the descriptive values of the facts so they are all called as a Dimensional Tables.

The dimension table is not denormalized. Start schema consists of fact table that are nothing but numerical values that are additive measurements and dimensional table contain smaller and descriptive data.

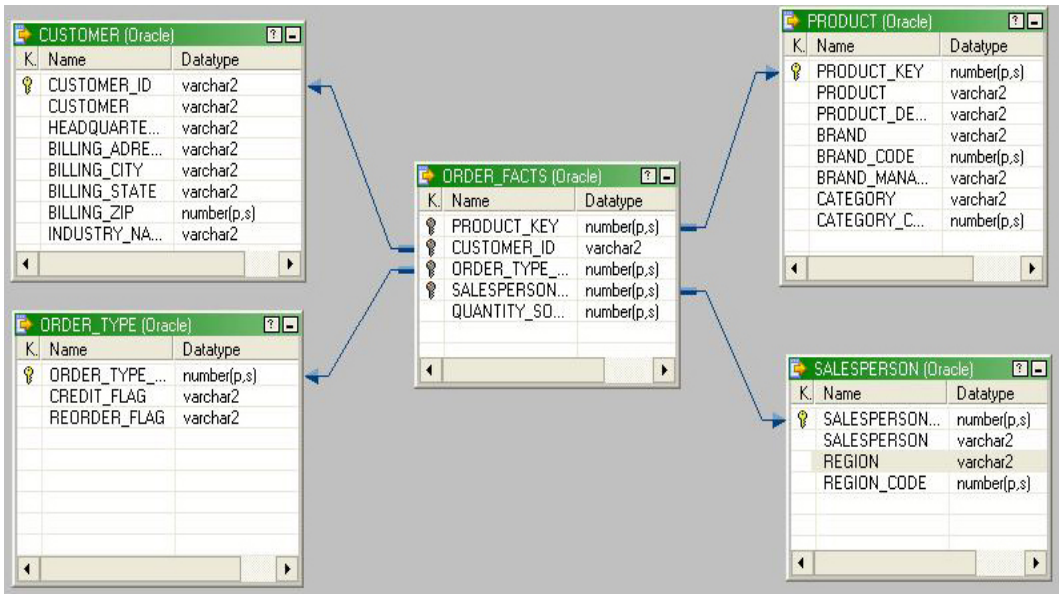



**Figure 9. Star Schema**

### 10.2.1. Objective

Identify the NULL values that are replacing with constant values for increasing the performance of the Data in the Data Warehouse Taking Source as PRODUCT table from the src_data.

### 10.2.2. Source

In this Source PRODUCT table we need to identify the null values in the CATEGORY_CODE column.

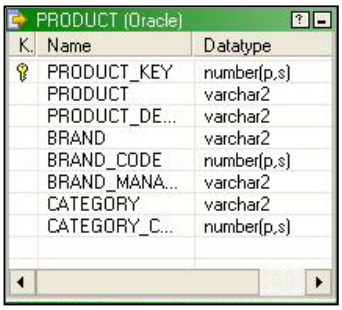



**Figure 11. Source Preview**

The Preview of the source data contains many NULL values in the PRODUCT table.

### 10.2.3. Target

First we need to identify the NULL values for this we need to create three target tables.

1) Identify the NULL values

2) Identify the NOT NULL values

3) Filling NULL values with constant Value.

T_NULL_PRODUCT table Contain only NULL Values. T_REJ_NULL_PRODUCT table contain NOT NULL Values. T_UPDATE_NULL_PRODUCT updating the null values which contain the UPDATED NULL Values.

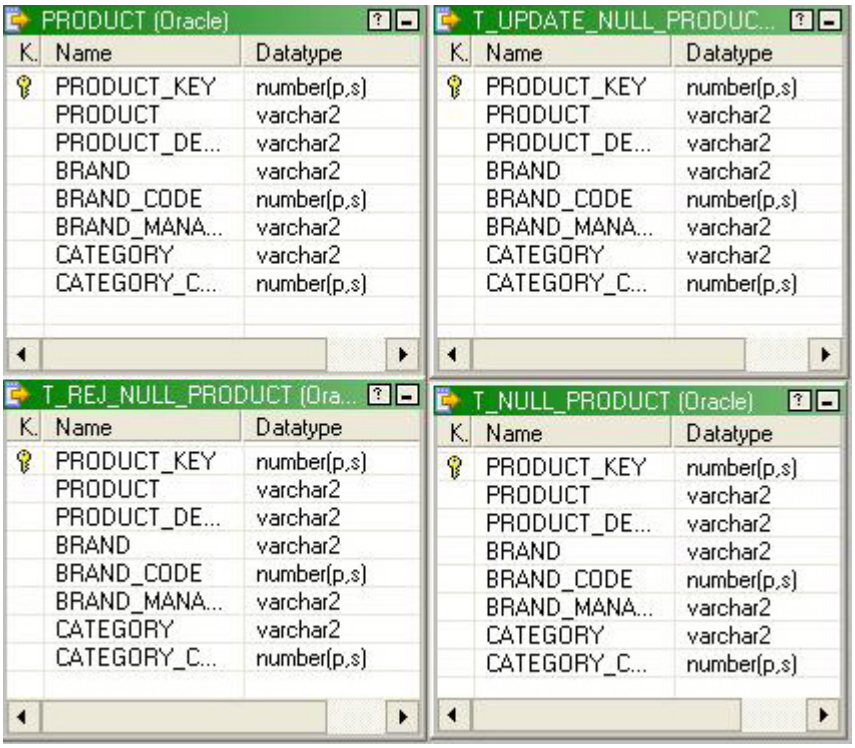



**Figure 12. T_UPDATE_NULL_PRODUCT, T_REJ_NULL_PRODUCT,T_NULL_PRODUCT**

### 10.2.4. Mapping

The Mapping Design used to identify the NULL values and also automatically it updates the NULL Values with constant Value '999' for improving data access or data performance in the Data Warehouse.

#### 10.2.4.1. EXP_NULL_REJ

In the expression Transformation we are creating two new ports: One for identifying the NULL values and another port for Not NULL values. Here we are going to develop two Expressions for the separating the NULL and NOT NULL Values.

#### 10.2.4.2. RTR_PRODUCT

The Router Transformation used to separating or distribute the NULL values into one Target Table **T_NULL_VALUES.** NOT NULL values into **TA_REJECT_EMP.** For this we are creating two groups one for NULL Values & another NOT NULL Values. In the second mapping we are using again expression transformation in that we again using the expression for updating the values into the TA_NULL_VALUES.

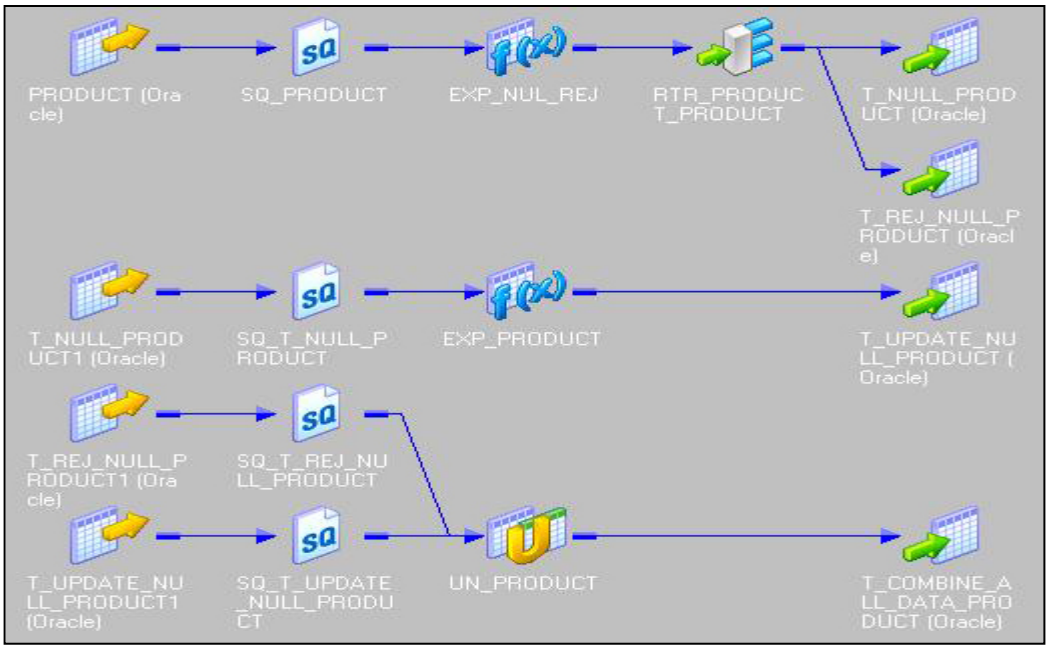



**Figure 13. Mapping for Replace Null Values with constant Values '999'**

| Port Name | Expression |
|---|---|
| **O_NULL_Values** | **IIF**(**ISNULL**(COMM),1,0) |
| **O_NOT_NULL_Values** | **IIF**(COMM != NULL, 1, 0 ) |

**Figure 14. Business Logic**

| Group Name | Expression |
|---|---|
| **O_NULL_Values** | **TR_VALUE = 1** |
| **O_NOT_NULL_Values** | **FA_NOTNULL = 1** |

**Figure 15. Business Logic2**

In the third mapping we are going to load complete data into single table.

| Port Name | Expression |
|---|---|
| **REMOV_SPACE** | **LTRIM**(**RTRIM**(COMM )) |
| **REPLACE_NULL** | **IIF( REMOV_SPACE, NULL,999)** |

**Figure 16. Business Logic3**

### 10.2.5. Target Result:

In the first Mapping we are going to identify the '**NULL'** Values and separate them into two target tables. One conatain complete **NULL** Values(**T_NULL_PRODUC)** , another Table Contain **NOT NULL** Values(**T_REJ_NULL_PRODUCT)**.

| PRODUCT_KEY | PRODUCT | PRODUCT_DESCRIPTION | BRAND | BRAND_CODE | BRAND_M... | CATEGORY | CATEGORY_CODE |
|---|---|---|---|---|---|---|---|
| 12345 | Appleipod | Etertainment | APPLE | 1234 | Palmer | home Entertainm... | NULL |
| 14567 | Engine | MaruthiEngine | Maruthi□ | 5123 | Ellender | Auto accessories | NULL |
| 19032 | Book | The love Dare | Alex | 6143 | Connie | books & Magazine... | NULL |
| 13564 | Camera | New model | Canano | 8061 | Rory | cameras & Optics | NULL |

**Figure 17. Null Values in PRODUCT TABLE**

Here is the second Target Table (**Preview of the Data in the Target Table) T_REJ_NULL_PRODUCT.**

The source **Product Table** Contain the NULL values we are separating that into two target table, After the we need to fill the NULL Values with a Constant Values, After we need to merge the two target Tables for loading the data into Data Warehouse. This process is completely done in Dynamical manner. The **NULL** Values are replaced by '999' constant value. The '999' Constant values in one table, and not NOT NULL Values in another table, so its very difficult to take the decision that's way we need to merge the data.

| PRODUCT_KEY | PRODUCT | PRODUCT_DE... | BRAND | BRAND_CODE | BRAND_MANAGER | CATEGORY | CATEGORY_CODE |
|---|---|---|---|---|---|---|---|
| 12345 | Appleipod | Etertainment | APPLE | 1234 | Palmer | home Entertain... | 999 |
| 14567 | Engine | MaruthiEngine | Maruthi□ | 5123 | Ellender | Auto accessories | 999 |
| 19032 | Book | The love Dare | Alex | 6143 | Connie | books & Magazi... | 999 |
| 13564 | Camera | New model | Canano | 8061 | Rory | camer Auto accessor | 999 |

**Figure 18. The NULL VALUES Replaced by a Constant Value '999'**

| PRODUCT_KEY | PRODUCT | PRODUCT_D... | BRAND | BRAND_CODE | BRAND_M... | CATEGORY | CATEGORY_CODE |
|---|---|---|---|---|---|---|---|
| 12345 | Appleipod | Etertainment | APPLE | 1234 | Palmer | home Ente... | 999 |
| 14567 | Engine | MaruthiEngine | Maruthi□ | 5123 | Ellender | Auto acce... | 999 |
| 19032 | Book | The love Dare | Alex | 6143 | Connie | books & M... | 999 |
| 13564 | Camera | New model | Canano | 8061 | Rory | cameras &... | 999 |
| 9876 | TV | telivision | sony | 567 | SMITH | electronic | 987 |
| 11691 | Appleipod | Etertainment | APPLE | 1234 | Palmer | home Ente... | 8061 |
| 12321 | Engine | MaruthiEngine | Maruthi | 5123 | Ellender | Auto acce... | 512 |
| 14343 | Book | The love Dare | Alex | 6143 | Connie | books & M... | 1143 |
| 15323 | Camera | New model | Canano | 8061 | Rory | cameras &... | 2190 |
| 93522 | car | New modelW | BMW | 512 | Cordell | cars & Bikes | 3430 |
| 84485 | greeting | Valentains day | Archies | 1143 | Scott | Charity | 4402 |
| 84345 | shirt | cotton shirt | Levis | 2190 | Porfirio | Clothing &... | 9165 |
| 99089 | Indian coins | old category | Indian | 3430 | Pablo | Coins & no... | 3652 |
| 12564 | cycle | jim cycling | Rofus | 4402 | Dewey | Fitness & ... | 9061 |
| 78990 | memory card | memory card | transend | 9165 | Sergio | Memory c... | 1012 |
| 45764 | mobile | samsung jim cycling | msung | 3652 | Bennie | Mobile Acc... | 1345 |
| 98833 | DvD | movies | Aditya | 9061 | Lamont | Movies & ... | 3254 |
| 45567 | Perfum | Jasmin & roses | Royal | 1012 | Humberto | ,Cosmetic... | 5678 |
| 49988 | shoes | sports wear | addidas | 1345 | Orval | Shoes | 3652 |
| 34521 | laptop | third generation | lenovo | 3254 | Myron | Computer ... | 9061 |
| 45789 | Toolkit | new model kit | Sandi | 5678 | Enoch | Tools & ha... | 1012 |
| 21563 | Teddy Bear | Smooth & cot... | Archies | 6789 | Jerald | Toys & Ga... | 1345 |
| 39872 | Hand watch | Wrist watch | Fastract | 7894 | Anton | Watches | 1245 |

**Figure 19. Complete Data Loaded into a Single Target Table**

After the combing to Table the total data available into a Single Table. (**T_COMBINE_ALL_DATA_PRODUCT).**

## 10.3. Non Duplication

It is the characteristics of having all unique values for the data fields.

### Applies to:

Informatica Power Center 8.6.0

**Summary**

This article describes a solution to identify duplicate and distinct records in a source.

**Table of Contents**



**10.3.1. Objective**

To segregate distinct and duplicate records in a source, solution described below uses an expression transformation. This solution exploits the concept that all input ports are evaluated first, then variable ports are evaluated and all output ports are evaluated in last.

**10.3.2. Source Definition**

Source is EMPLOYEE table as shown in figure. In this table we didn’t implement the any primary key that the reason it have the duplicate values of every employees, if any employee details are duplicated we are going to eliminate that rows and maintain the unique values in the source table.

EMPLOYEE (Oracle)

| K. | Name | Datatype | Length/Precision |
|---|---|---|---|
| | EMPNO | number(p,s) | 4 |
| | ENAME | varchar2 | 10 |
| | JOB | varchar2 | 9 |
| | MGR | number(p,s) | 4 |
| | HIREDATE | date | 19 |
| | SAL | number(p,s) | 7 |
| | COMM | number(p,s) | 7 |
| | DEPTNO | number(p,s) | 2 |

**Figure 20. Relational Source EMPLOYEE**

### 10.3.3. Target Definition

There are two targets in sample mapping. All duplicate records found in source table will be inserted into target table T_DUPLICATE_EMP and distinct records will be inserted into target table T_DISTINCT_EMP.

### 10.3.4. Mapping

Figure shows the mapping designed to identify duplicate and distinct records.

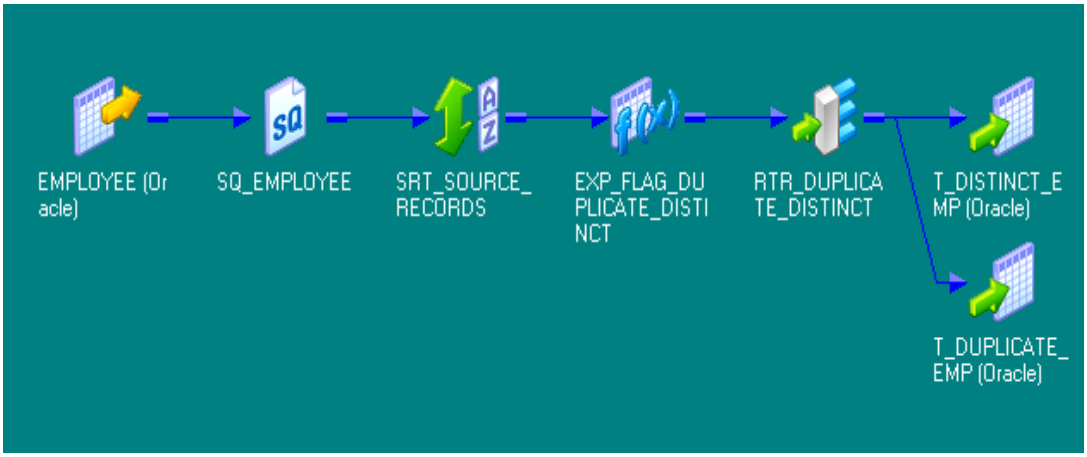



**Figure 21. Mapping Design**

#### 10.3.4.1 Transformations

**SRT_SOURCE_RECORDS**

Sorter transformation is used after Source Qualifier to sort source records. All ports are selected as key.

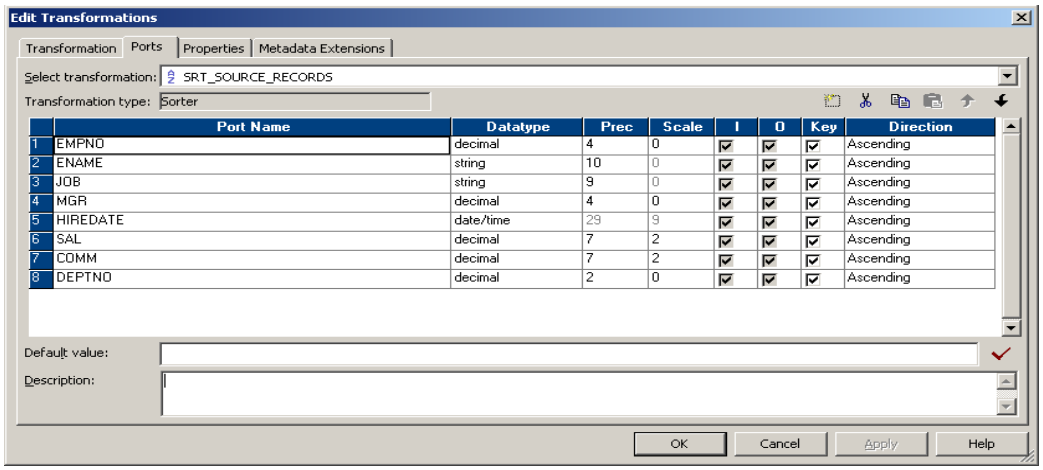



**Figure 22. Sorter Transformation**

## EXP_FLAG_DUPLICATE_DISTINCT

Expression transformation is used to flag a record as either duplicate or distinct. Current record (v_CurrentRec) is compared with previous record (v_PrevRec) and if records match v_IsDuplicate variable is set to 'Y' else it is set to 'N'.

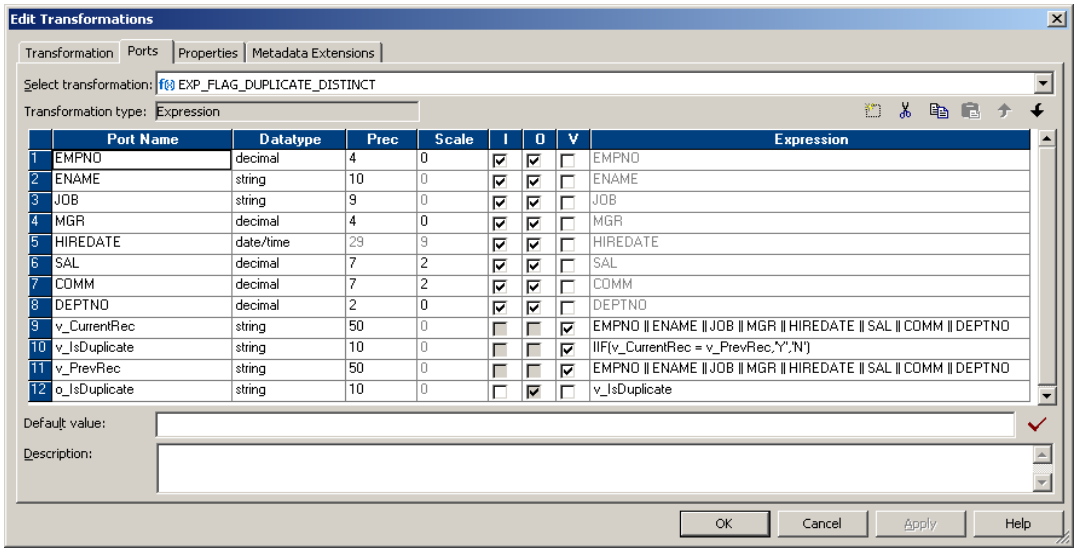



**Figure 23. Expression Transformation**

Table shows expression used for variable and output ports

**Table. Logic used in Expression Transformation**

| Port | Name Expression |
|---|---|
| **V_CurrentRec** | EMPNO\|\|ENAME\|\|JOB\|\|MGR\|\|HIREDATE\|\|SAL\|\|COMM\|\|DEPTNO |
| **V_ISDuplicate** | IIF(V_CURRENTREC=V_PREVREC,'Y','N') |
| **V_PrevRec** | EMPNO\|\|ENAME\|\|JOB\|\|MGR\|\|HIREDATE\|\|SAL\|\|COMM\|\|DEPTNO |
| **O_ISDuplicate** | V_ISDuplicate |

| EMPNO | ENAME | JOB | SAL |
|---|---|---|---|
| 7782 | CLARK | MANAGER | 2450 |
| 7844 | TURNER | SALESMAN | 1500 |
| 7369 | SMITH | CLERK | 800 |
| 7654 | MARTIN | SALESMAN | 1250 |
| 7788 | SCOTT | ANALYST | 3000 |
| 7369 | smith | salesman | 9000 |
| 7566 | JONES | MANAGER | 2975 |
| 7902 | FORD | ANALYST | 3000 |
| 7934 | MILLER | CLERK | 1300 |
| 7698 | BLAKE | MANAGER | 2850 |
| 7876 | ADAMS | CLERK | 1100 |

| EMPNO | ENAME | JOB | SAL |
|---|---|---|---|
| 7499 | ALLEN | SALESMAN | 1600 |
| 7521 | WARD | SALESMAN | 1250 |
| 7839 | KING | PRESIDENT | 5000 |
| 7900 | JAMES | CLERK | 950 |

**Figure 24. Unique Records**

## 11. Conclusion

ETL process is an integral component of data warehousing environment. The main contribution of this paper is in the development of Data Warehouse which gives quality Data Warehouse. The model defines Several Quality factors to improve the performance of the data warehouse. We have develop the own ETL scenarios for implementing the framework for the taking the data from the different source. Next design the ETL work flows, after we can design the logical model for this, and then mapping from sources to logical model. Finally we present the data in the Data Warehouse to increase the performance the system and which is best model to improve the performance and fast data access for data warehouse. This work is enhanced to Reporting work.

## References

[1] P. Pahawa, S. Taneja and G. Thaur, "UCLEAN: A Requirement based Object Oriented ETL Frame Work", IJCSES, **(2011)**.
[2] P. Kiran, S. Satish Kumar and N. P. Kavya, "Modeling Extraction Transformation Load Embedding privacy preservation using UML", International journal of computer application, vol. 50, no. 6, **(2012)** July.
[3] A. Cembalo, F. M. Pisano and G. romano, "An approach to document Warehousing System Lifecycle from Textual ETL to Multidimensional Queries", 6th international conference on complex, Intelligent, and Software Intensive Systems.
[4] M. M. Hamad and A. Abdulkhar Jihad, "An Enhanced Technique to Clean Data in DWH", Development in E-systems Engineering, **(2011)**.
[5] 24/7 Real- Time Data Warehousing: A Tool for Continuous Actionable Knowledge-IEEE Annual Computer software and Application Conference.
[6] The business intelligence- based integrated information management system in the road transport research-ICIMIMIE, **(2012)**.
[7] A. Bustamanta Martinez and E. Amaru Galvis-Lista, "Modeling Techniques for Extraction Transformation and Load Process".
[8] F. Nur Savitri and H. Laksmiwati, "Study of Localized Data Cleaning Process for ETL Performance Improvement in Independent Datamart", ICEEI, **(2011)**.
[9] "Data warehouse as a Backbone for Business intelligence L Issues and Challenges", European journal of Economics, finance and administrative science ISSN 1450-2275, **(2011)**.
[10] S. Ghosh, S. Goswami and A. chakrabarti, "Outlier detection from ETL Execution Trace".
[11] V. Sagar, "IDSS: An object process based DWH construction Method", **(2011)**.
[12] Application research of DWH technology in decision- making of drug distribution enterprise, IEEE, **(2010)**.
[13] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Second Edition, pp. 106-200.
[14] S. Ghosh, "Outlier detection from ETL Execution trace", IEEE, **(2011)**.
[15] Informatica University books: Getting Started 27-60. Power center Designer 30-85, 95-120.
[16] Star/snow-flake schema Driven Object-Relational DWH Design and Query Procession Strategies, UGC Research Grants Council of HKSR under grant733/96E.
[17] M. Levene and G. Loizou, "Why is the snowflake schema a Good DWH Design?", School of computer Science and information System, Birkbech College University, **(2010)**.

[18] A project-oriented data warehouse for construction Thammasak Rujirayanyong a, Jonathan J. Shi b, a Accepted 16 November 2005, Elsevier, **(2010)**.
[19] ORACLE Series / Oracle8i Data Warehousing / Corey, Abbey, Abramson, Taub / 2675-2 / Chapter 7
[20] Modeling and managing ETL processes Alkis Simitsis National Technical University of Athens, Dept. of Electrical and Computer Eng., Computer Science Division, Iroon Polytechniou 9, Zografou 15773.
[21] Informatica University books: Workflow Administration, pp. 235-265.

## Authors

**Kushanoor Akbar**, JNTUA, Anantapur university, Madanapalle Institute of Technology & Science, Kadiri Road Angallu (Vill), Madanapalle-517 325, India .E-Mail: akbar.akbar8@gmail.com.

**Dr.S.Murali Krishna B.Tech, M.Tech, Ph.D**, JNTUA, Anantapur university, Professor and Head in CSE Dept, Madanapalle Institute of Technology & Science, Kadiri Road Angallu (Vill), Madanapalle-517 325, India. E-mail: drmuralimits@gmail.com

**T. Vidya Sagar Reddy B.Tech**, **M.Tech**., Senior Data Warehouse consultant, Capgemini PVT LTD, Bangalore, India. E-Mail: vidyaveda4@gmail.com.