

O&T Dashboard

ETL High Level Design Document

**ScheduAllPost**

DOCUMENT CONTROL

|  |  |  |  |
| --- | --- | --- | --- |
| **Prepared by** | **Role** | **Date of Preparation** | |
| Rajasekhar B | ETL Lead | 3/19/2019 | |
| **Reviewed by** | **Role** | **Date of Review** | |
| PaulNaveen | ETL Architect | 3/20/2019 | |
| **Reviewed by** | **Role** | **Date of Review** | |
| Anthony Mickey | Business Analytics |  | |
| **Approved by** | **Role** | **Date of Approval** | |
| Rajasekar Akula | Project Manager |  | |
| **Version Description** | Initial Version | **Version #** | 1.0 |

REVISION History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S No** | **Date** | **Version #** | **Section Changed** | **Details of changes made** | **Approved by** |
| 1 | 3/19/19 | 1.0 | ALL | Added details in all sections |  |

**Disclaimer – The Information in this document is subject to change and must be reviewed frequently**

Table of contents

[1. INTRODUCTION 4](#_Toc516581536)

[1.1 Overview 4](#_Toc516581537)

[1.2 Solution Approach 4](#_Toc516581538)

[1.3 Technology Stack 4](#_Toc516581539)

[2. Etl Process Flow](#_Toc516581540) 4

[2.1 Sources Information](#_Toc516581541) 4

[2.2 Target Information 5](#_Toc516581542)

[2.3 ETL Architecture 5](#_Toc516581543)

[2.4 ETL Process Flow 5](#_Toc516581544)

[2.5 Extraction Strategy 7](#_Toc516581545)

[2.6 DATA Cleansing 7](#_Toc516581546)

[2.7 TRANSFORMATIONS 7](#_Toc516581547)

[2.8 LOAD Strategy 8](#_Toc516581548)

[2.9 Metadata Management 9](#_Toc516581549)

[2.10 Reconciliation 9](#_Toc516581550)

[2.11 Exception And Error Handling 9](#_Toc516581551)

[2.12 Restartability Strategy 9](#_Toc516581552)

[2.13 Alerts And Notification Strategy 9](#_Toc516581553)

[2.14 Reprocessing Of Bad Data 9](#_Toc516581554)

[2.15 Archiving And Purging Strategy 9](#_Toc516581555)

[3. Production Support Process](#_Toc516581556) 9

[4. APPENDIX 9](#_Toc516581557)

[4.1 ETL Data LLD 9](#_Toc516581558)

[4.2 OTHER Settings & Parameters 1](#_Toc516581559)0

## 1. INTRODUCTION

## 1.1 Overview

The overall objective is to integrate ScheduAllPost data from ScheduALL DB into OT data warehouse environment. The data will be used by theCenter portal for business reporting needs.

## 1.2 Solution Approach

The diagram below explains the Solution Approach for the ScheduAllPost data integration into OT environment.

## 1363538344_Web_Database 1363538911_iDatabase 1363538911_iDatabase

Tables/Informatica

Informatica

MSSQL DB TD Stage and Dim Tables TD Target Tables

The ETL team will extract data from ScheduAllPost DB hourly and load the same to Teradata staging table then to target dimension and dim/fact tables.

## 1.3 Technology Stack

|  |  |
| --- | --- |
| Technology | S/W & Hardware |
| ETL tool | Informatica 10.1 |
| Connectivity | ODBC, TPT API |
| Scheduler | Control-M |
| Source Server | mtddbsched501.tfayd.com\100.115.104.58,1433 |
| Target Server | tdprodla.tfayd.com |
| Source Database | Schedwin |
| Target Database | OTSDW\_VM, OTSDW\_DBO, OTSDW\_STAGE (Teradata) |

## 

## 2. ETL Process Flow

ETL process flow is mentioned below

## 2.1 Sources Information

|  |  |
| --- | --- |
| Source System | MSSQL |
| Data Provided | Yes |
| Server | mtddbsched501.tfayd.com\100.115.104.58,1433 |
| Database | Schedwin |
| Connectivity | ODBC |
| Average Daily Volume | 5000 |

## 2.2 Target Information

Target database information is mentioned below

|  |  |
| --- | --- |
| Target System | Teradata |
| Server | tdprodla.tfayd.com |
| Database | OTSDW\_DBO, OTSDW \_STAGE, OTSDW \_VM |
| Connectivity | ODBC. TPT API |

## 2.3 ETL Architecture

The diagram below explains the ETL architecture followed for ScheduAllPost data integration into OT environment.

|  |
| --- |
| Support Inbox  OTS DATAMART  Stage Table  Dimension Tables  VMBI layer  Informatica (Full refresh ETL)  Fact Tables  On time / Delay process  Data Reconciliation  Reports  ScheduAllPost tables    STAGE TABLES  STAGE TABLES |

## 2.4 ETL Process Flow

The diagram below explains the ETL process flow followed for ScheduAllPost DB data integration into OT environment.

|  |
| --- |
| ScheduleAllPost DATAMART  Reporting  Stage Table  Dimensions Fact Tables  VMBI Layer    ETL using Informatica  ScheduAllPost |

2.4.1 Staging Area Load

Truncate and load the stage table. For STG\_WO,STG\_QUSAGE,STG\_QUOTES and STG\_SEVT tables we are maintain 15 days’ worth of data for debugging any issues in future as we don’t have access to source tables

Find the STAGE tables structure in the following attachment.

**

2.4.2 Fact Load

For SPOST\_DIM\_QUOTE and SPOST\_DIM\_WORK\_ORDER we are using SCD type II whereas for others its SCD type I

Find the Dimension table structure in the following attachment.



2.4.3 Fact Load

We are using SCD type II to load the fact tables. Find the Dimension table structure in the following attachment.



## 2.5 Extraction Strategy

We are extracting ScheduAllPost source data daily from Database. Here we are loading the source data completely to stage table. Before loading we are truncating the stage table. Once stage load completes, we will load the data from stage table to respective target tables.

## 2.6 DATA Cleansing

Detailed data cleansing and transformation document has been attached below.

## 2.7 TRANSFORMATIONS

Details about transformations will be provided in detail in ETL low level design document.

## 2.8 Load Strategy

Loading Stage Table:

Stage tables is truncate load and dimension tables are loaded UPSERT and SCD type II (SPOST\_DIM\_QUOTE and SPOST\_DIM\_WORK\_ORDER ) and fact tables are SCD type II

2.8.1 Incremental Load

We are pulling incremental data for the tables which has LAST\_MOD column

2.8.2 Historical Load

If we want to perform historical load then we need to edit the parameter file to change the LAST\_MOD columns

2.8.3 Load Order

The load order for ScheduAllPost is mentioned below.

|  |  |  |  |
| --- | --- | --- | --- |
| SL. NO. | ETL PROCESS NAME | DEPENDENT ON | LOAD ORDER |
| 1 | OTS\_SPOST\_STG\_load.ksh | NA | 1 |
| 2 | OTS\_SPOST\_DIM\_load.ksh | Dependent on STAGE load | 2 |
| 3 | OTS\_SPOST\_FCT\_load.ksh | Dependent on STAGE and DIM load | 3 |

2.8.4 Control-M Scheduling

The job schedule order for ScheduAllPost daily is mentioned below

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SL. NO. | JOB NAME | UNIX SCRIPT | START TIME | PREDECESSOR | SUCCESSOR |
| 1 | STMDPSTG | OTS\_SPOST\_STG\_load.ksh | 06:00 AM EST | - | - |
| 2 | STMDPDIM | OTS\_SPOST\_DIM\_load.ksh | - | STMDPSTG | - |
| 3 | STMDPFCT | OTS\_SPOST\_FCT\_load.ksh | - | - | STMDPFCT |

## 2.9 Meta Data Management

The Metadata Management procedures are as follows:

1. The management of the O&T metadata, audit fields and ETL process statistics in **EACS** tables.
2. **EACSNDW\_VM.FACT\_JOB\_RUN** Table which is used for job load statistics.
3. **FACT\_DATA\_QUALITY\_SUMMARY** Table which is used for data validation.

## RECONCILIATION

The data reconciliation procedures are as follows:

1. Validating Source File vs Stage table record counts.
2. Validating Stage Table vs Fact Table record counts as and when required.

## 2.11 Exceptions and Error Handling

After all the jobs get completed successfully, a success mail will be sent to the respective groups stating that the load completed successfully. In case of any failure in the jobs, failure notification mail will be sending to respective groups and the issue should be fixed and restart the jobs.

## 2.12 Restartability Strategy

The table below gives the Restartability strategy to be used during the failures.

|  |  |
| --- | --- |
| Process | Action |
| Scheduler fails | Contact Support Team. |
| Try to run batch job manually. |
| Network failure | Cancel current running and all dependent jobs |
| Contact Network Admin group |
| Restart the scheduler and verify |
| Server Failure | Cancel current running and all dependent jobs |
| Contact Infrastructure Admin group |
| Restart the scheduler and verify |
| Database failure | Cancel current running and all dependent jobs |
| Contact DBA group |
| Restart the scheduler and verify |

## 

## 2.13 Alerts and Notification Strategy

The notification e-mail will be sent to application support on success or failure of each run. These e-mails will contain the following information:

1. Details for the build run: for example; (number of records loaded in the target - number of records rejected - and time taken to run the build).
2. On every successful completion of ETL process, an ETL statistics audit report will be attached with the notification email.

|  |  |  |  |
| --- | --- | --- | --- |
| Stage to Notify | What to Notify | Mechanism | User Group to Notify |
| Start of Extraction | ETL Started | E-Mail | ontssupport@nbcuni.com |
| Failure of Task | Task Failed | E-Mail | ontssupport@nbcuni.com |
| Completion of ETL | ETL Completed | E-Mail | ontssupport@nbcuni.com |
| Reconciliation | Reconciled | E-Mail | ontssupport@nbcuni.com |

## 2.14 REPROCESSING OF BAD DATA

Procedures to reprocess Bad Data are as follows:

1. An automated internal email within our team.
2. Team will review the data.
3. Team will correct any issues in the code.
4. Team will reprocess the data.

## 2.15 ARCHIVING & PURGING STRATEGY

The Archive and Purging procedures occur after each successful ETL process completion and are as follows:

1. Source file is archived to [Informatica batch server archive file directory].
2. Data feeds for the past 30 days are archived.
3. The archived files can be traced back to 12 months.

## 3. Production Support Process

The job schedule order for OT ScheduAllPost daily is mentioned below

|  |  |
| --- | --- |
| Category | Comments |
| Source file issues | Support Team to connect business and resolve issue. |
| Informatica job failure | As soon as support team get email notification. We will start working on that issue. |
| Teradata issue | OT Support team work with Teradata team and resolve the issue. |
|  |  |

## 4. APPENDIX

## 4.1 ETL Data LLD

Detailed data cleansing and transformation logic has been given in the following LLD document.



## 4.2 Other Settings & Parameters

List of ScheduAllPost UNIX job scripts are listed in the attached document.



Details of ScheduAllPost Unix batch server directory structure is attached in the following document.