**Operation Manual**

**Data Ingestion to EDL Staging using IBM DIP**

*Ascena Brands*



Table of Contents

[1 Document Revision History 4](#_Toc22559152)

[2 Program Overview 5](#_Toc22559153)

[2.1 Overview 5](#_Toc22559154)

[2.2 Overview of Functionality and Key Processes 5](#_Toc22559155)

[2.2.1 DIP Ingestion Process 5](#_Toc22559156)

[2.3 Contributing Applications and Services 6](#_Toc22559157)

[2.4 Scheduled Events 6](#_Toc22559158)

[2.5 Infrastructure and Design 7](#_Toc22559159)

[2.5.1 Process Flow Diagram 7](#_Toc22559160)

[2.5.2 TWS Orchestration 8](#_Toc22559161)

[3 Security and Access Control 10](#_Toc22559162)

[4 Program Configuration 10](#_Toc22559163)

[5 Configuration Management 11](#_Toc22559164)

[5.1 Code Repository 11](#_Toc22559165)

[5.2 Backup and Restore Procedures 11](#_Toc22559166)

[6 Monitoring and Alerting 11](#_Toc22559167)

[6.1 Daily Operational Monitoring and Alerting 11](#_Toc22559168)

[6.2 Error Messages 11](#_Toc22559169)

[6.3 Health Checks 11](#_Toc22559170)

[7 Operational Tasks 11](#_Toc22559171)

[7.1 Deployment 11](#_Toc22559172)

[8 Failure and Recovery Procedures 12](#_Toc22559173)

[8.1 Troubleshooting 12](#_Toc22559174)

[9 Contact Details 12](#_Toc22559175)

[10 Additional Document References 12](#_Toc22559176)

[11 Appendix 13](#_Toc22559177)

Table of Figures

[Figure 1 - Scheduled Events Example 6](#_Toc19630299)

[Figure 2 - Process Flow Diagram - Data Ingestion to EDL Staging with IBM DIP(End-to-End) 7](#_Toc19630300)

[Figure 3 - Process Flow - DIP Processing 8](#_Toc19630301)

[Figure 4 - File and Data Orchestration - Job Stream](#_Toc19630307) 8

[Figure 5 - File and Data Orchestration - Decrypt Job 9](#_Toc19630308)

[Figure 6 - File and Data Orchestration - Start Job 9](#_Toc19630309)

[Figure 7 - File and Data Orchestration – Main Job 10](#_Toc19630310)

[Figure 8 - File and Data Orchestration – End Job 10](#_Toc19630311)

Table of Tables

Table 1 - Configuration Options 11

# Document Revision History

| Date | Author | **Revision Description** |
| --- | --- | --- |
| 2019/10/15 | George Brown | Initial Version |
| 2019/11/18 | George Brown | Corrected link to “Data Ingestion to EDL Staging with IBM DIP - TWS-Job\_Schedule-Ascena Retail.xlsx” |
|  |  |  |
|  |  |  |

# Program Overview

## Overview

|  |  |
| --- | --- |
| **Purpose** | All data feeds transmitted by EFT jobs into EDL(Enterprise Data Lake) edge node are ingested into staging layer using IBM DIP process.  Each data file from edge node travels through a sequence of feed specific processes/scripts(Decrypt Job, Start Job, Main Job and End Job) as part of DIP.  Once the data are loaded into the Staging layer, they may be processed into Conform Layer or Data Mart based on the type of data (those downstream processes are NOT covered in this document). |
| **Document Scope** | This operation manual will contain documentation regarding the core components of Data Ingestion to EDL Staging with IBM DIP process (all brands). Documentation on proceeding or succeeding processes or programs will not be included in this operation manual. However, proceeding or succeeding documentation may be referenced. |

## Overview of Functionality and Key Processes

### DIP Ingestion Process

Once the data files are landed in EDL edge node, they would be processed as per the TWS schedule for each file. Each file will have its corresponding job stream scheduled in TWS. In each job stream, the files would undergo the below list of processes(in the given sequence).

1. Decrypt Job - The files placed in EDL edge node by EFT jobs were encrypted by IBM public key. This decrypt process would decrypt the files using IBM private key.
2. Start Job - In this process, files undergo duplication check and conversion like fixed file to delimiter file(as needed). Then this process increments the batch id in the batch id control table based on the file drop schedule(daily, monthly etc.)
3. Main Job - This is where the data from the files are loaded into corresponding Staging Layer tables after submitted to quality validation. Once the data is loaded into Staging tables, the files are moved to Archive direcotories in edge node and HDFS.
4. End Job - After the data load is completed, this process would update batch Id control table to signify the completion of the load. It also does the final clean up activities(cleaning up the files after loading and archiving).

## Contributing Applications and Services

Data ingestion to EDL staging with IBM DIP use the following applications and services.

* TWS - Each file loaded into EDL edge node has a corresponding Job stream for IBM DIP process which would execute Decrypt Job, Start Job, Main Job and End job for the particular file ID. These job streams are scheduled through TWS scheduler based on the file frequency.
* HDFS - Hadoop Distributed File System is where data are stored physically for the Hive tables. Once the Staging Layer load is done, raw files are alos being archived in HDFS.
* MySQL - MySQL database is used as the Metastore for Hive tables.
* HIVE - Raw Layer/Staging Layer tables are created and store in Hive Database.
* Encryption Keys - IBM private keys are used in the DIP process to decrypt the data files that are encrypted using IBM public keys by the EFT jobs.

## Scheduled Events

The spreadsheet “Data Ingestion to EDL Staging with IBM DIP - TWS-Job\_Schedule-Ascena Retail.xlsx”(reference is given later in the Additional Documents Reference section) will provide schedules for all jobs and events by file ID. Below is a sample screenshot showing the schedule for each file ID. Below times are in CST.

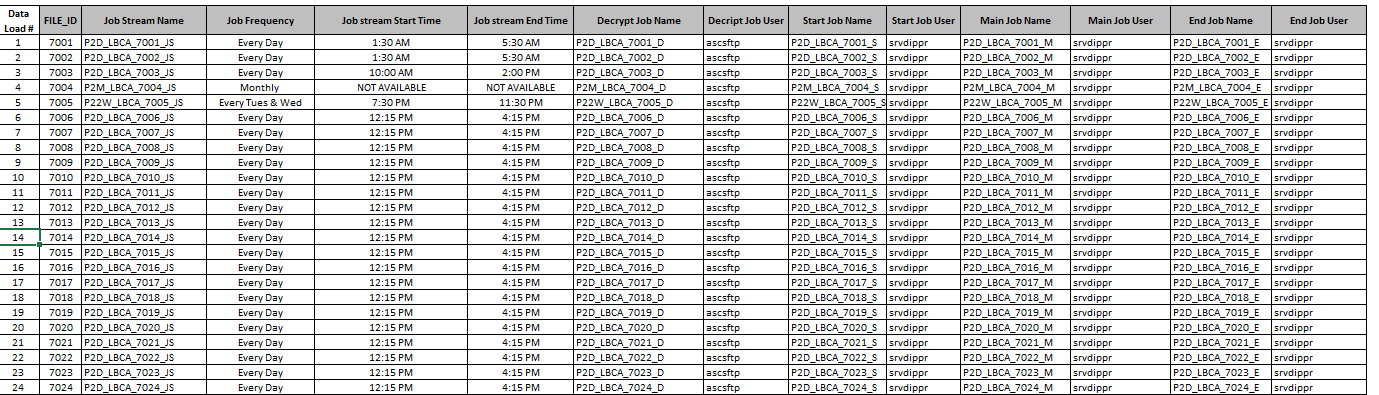


Figure 1 - Scheduled Events Example

## Infrastructure and Design

### Process Flow Diagram

The below diagram shows the steps a file in the EDL edge node takes in order to get ingested into the staging database in Hive with IBM DIP process.

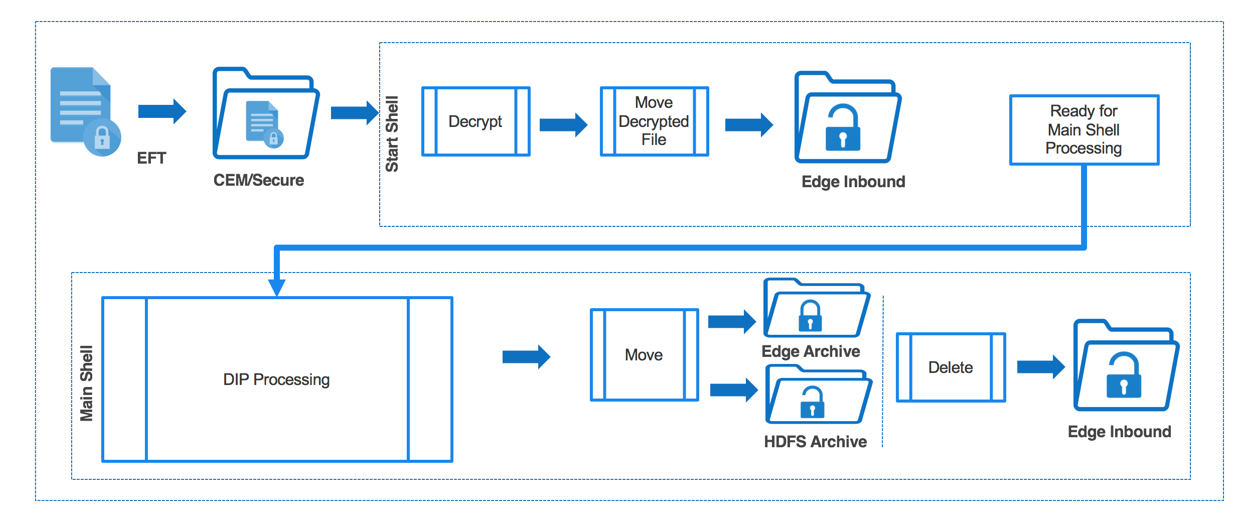


Figure 2 - Process Flow Diagram – Data Ingestion to EDL Staging with IBM DIP(End-to-End)



Figure 3 - Process Flow – DIP Processing

### TWS Orchestration

The “Data Ingestion to EDL Staging with IBM DIP - TWS-Job\_Schedule-Ascena Retail.xlsx” spreadsheet contains the detailed list of source files and their associated Job Stream, Descrypt job, Start Job, Main Job and End job. Details of these jobs are depicted below.

#### Job Stream

Each file ID has a corresponding job stream that’s scheduled in TWS. Below screenshot shows the file details and corresponding Job stream names. The job stream name is derived as “<Environment><Edge Node#><Frequency>”\_”<Brand/Brand Group>”\_”<File ID>”\_””JS””.



Figure 4 - File and Data Orchestration – Job Stream

#### Decrypt Job

Below screenshot shows how each file landed in EDL edge node is being decrypted. The decrypt process would run only if the script finds a “.ctl” file in the corresponding folder for a specific file in “secure” directory as given below.



Figure 5 - File and Data Orchestration – Decrypt Job

#### Start Job

Once the decrypt process is done, each file would undergo a sequence of processes scripted in Start Job. There are 8 parameters being passed in the Start Job script viz Home Directory, File ID, Table ID, Target File path, Archive File path, Secure File path, source file name and fixed/delimited flag. In this job/step, each file would be validated for duplication(if the same version of the file was already processed) and converted to delimited if it was a fixed file(as needed). Also, it updates the batch ID in batch ID control table and moves the data and control files from secure to inbound folder.



Figure 6 - File and Data Orchestration – Start Job

#### Main Job

Once the Start job is completed, Main Job script for the corresponding file ID is getting triggered. There are 12 parameters being passed in the Main Job script viz Home directory, File ID, View ID, Table ID, Load Type, Exec Engine, Source file path, Secure file path, Target file path, HDFS file path, HDFS archive file path, Header-footer flag. The significant parameters corresponding to each file are given in the spreadsheet(sample below).

This script performs data validation(row count, delimiter check etc), loading data into HDFS,loading data into Hive tables based on load type and archives the files into HDFS archive folder and edge node archive folder once the data is loaded/ingested into Staging layer tables.



Figure 7 - File and Data Orchestration – Main Job

#### End Job

Once the data has been ingested into Staging layer tables, End Job script corresponding to the data file gets triggered. There are two parameters being passed in the End Job script viz, Home directory and File ID.

In this step, the current and last batch\_id is being updated in the batch id control table followed by final clean up activities.



Figure 8 - File and Data Orchestration – End Job

# Security and Access Control

* + - The decrypt job/script should run using “ascsftp” user.
    - Start job, Main job and End job scripts should run using “srvdippr” user.

# Program Configuration

Data ingestion to EDL Staging can be configured/reconfigured in the following aspects.

| Configuration Option | Method to Configure |
| --- | --- |
| Metadata Changes (feed changes) | Complete metadata spreadsheet change, create LOPS to IBM, share the updated Metadata sheet with IBM . |
| TWS Job Schedule Changes | LOPS to IBM, Request updated TWS spreadsheet after change |
|  |  |
|  |  |

Table 1 - Configuration Options

# Configuration Management

## Code Repository

* IBM manages all the DIP scripts and Jobs

## Backup and Restore Procedures

* IBM manages the backup and restore procedures

# Monitoring and Alerting

## Daily Operational Monitoring and Alerting

* IBM manages

## Error Messages

* TWS Jobs – IBM supports the TWS jobs for DIP processes currently.

## Health Checks

There are no known health checks currently in place.

# Operational Tasks

## Deployment

* Metadata changes – Update the metadata accordingly, share it with IBM support. Create LOPS ticket and get the estimation from IBM. Create CR in ivanti, get it approved in GCAB and share the implementation plan with IBM. Coordinate the changes on the deployment day.
* TWS schedule changes – Create LOPS to IBM team and share the TWS schedule changes required. Create CR in ivanti, get it approved in GCAB and share the implementation plan with IBM. Coordinate the implementation.

# Failure and Recovery Procedures

## Troubleshooting

* TWS Jobs – IBM team to triage the issue and provide fixes. Any changes to the process should go through normal Ascena change management process

# Contact Details

| Contact | Role | Email | Phone |
| --- | --- | --- | --- |
| IBM Prod Support Team | Maintains and monitors the DIP process | [ascprod@in.ibm.com](mailto:ascprod@in.ibm.com) | N/A |

# Additional Document References

| # | Document Name | Document Link | Notes |
| --- | --- | --- | --- |
| 1. | EFT Transmission to Edge Node - File and Data Orchestration.xlsx | [http://epm01/sites/IT\_Sites/EnterprisData\_Lake\_Operations/Shared%20Documents/Operation%20Manuals/EFT%20Transmission%20to%20Edge%20Node%20-%20File%20and%20Data%20Orchestration.xlsx](http://epm01/sites/IT_Sites/EnterprisData_Lake_Operations/Shared%20Documents/Operation%20Manuals/EFT%20Transmission%20to%20Edge%20Node%20Documents/EFT%20Transmission%20to%20Edge%20Node%20-%20File%20and%20Data%20Orchestration.xlsx) | This spreadsheet contains a list of ALL files transmitted to the EDL edge node along with details on the EFT jobs, Control-M jobs, Hygeine processes, and TWS jobs associated. |
| 2. | Data Ingestion to EDL Staging with IBM DIP - TWS-Job\_Schedule-Ascena Retail.xlsx | <http://epm01/sites/IT_Sites/EnterprisData_Lake_Operations/Shared%20Documents/Operation%20Manuals/Data%20Ingestion%20to%20EDL%20Staging%20with%20IBM%20DIP%20Documents/Data%20Ingestion%20to%20EDL%20Staging%20with%20IBM%20DIP%20-%20TWS-Job_Schedule-Ascena%20Retail.xlsx> | List of TWS jobs. |
| 3. | IBM - Operations Manual Document.docx | [http://epm01/sites/IT\_Sites/EnterprisData\_Lake\_Operations/Shared%20Documents/Operation%20Manuals/Reviewed%20and%20Approved/IBM%20-%20Operations%20Manual%20Document.docx](http://epm01/sites/IT_Sites/EnterprisData_Lake_Operations/Shared%20Documents/Operation%20Manuals/Data%20Ingestion%20to%20EDL%20Staging%20with%20IBM%20DIP%20Documents/IBM%20-%20Operations%20Manual%20Document.docx) |  |
| 4. | IBM Reqs Spreadsheet - Format(with sample data).xlsx | <http://epm01/sites/IT_Sites/EnterprisData_Lake_Operations/Shared%20Documents/Operation%20Manuals/IBM%20Reqs%20Spreadsheet%20-%20Format(with%20sample%20data).xlsx> | Metadata sheet model |
| 5. | Data ingestion to EDL Staging with IBM DIP - Process flow.vsdx | <http://epm01/sites/IT_Sites/EnterprisData_Lake_Operations/Shared%20Documents/Operation%20Manuals/Data%20ingestion%20to%20EDL%20Staging%20with%20IBM%20DIP%20-%20Process%20flow.vsdx> |  |
| 6. |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Appendix