

# COMMONWEALTH OF PENNSYLVANIA

## DEPARTMENT OF HUMAN SERVICES

### INFORMATION TECHNOLOGY STANDARD

Name Of Standard: <b>Data Warehouse Access, Extraction, Transformation, Load, Deployment, and Scheduler Standards</b>	Number: <b>STD-EKMS009</b>
Domain: <b>Knowledge Management</b>	Category: <b>Data Warehouse</b>
Date Issued: <b>8/29/2008</b>	<b>DHS Bureau of Information Systems</b>
Date Revised: <b>2/19/2016</b>	

**Abstract:**

The purpose of this Standard is to establish enterprise-wide stands and guidance for using the Department of Human Service’s standard Extraction, Transformation, and Loading (ETL) tool, Informatica PowerCenter for Enterprise Data Warehouse (EDW) development.

ETL tools are used to extract data from the source file, transform the data into the proper formats, check the quality of the data, and load the data into the EDW.

**Technology Components:**

The mandatory toolset for ETL processing is Informatica PowerCenter.

**General:**

This standard applies to all developers, both commonwealth employees and business partners, who develop for the EDW for DHS. This policy ensures that all developed and implemented applications will facilitate enterprise-wide interoperability and standardization. Any questions or issues pertaining to EDW Informatica standards should be directed to the EKMS Informatica Team which may be contacted via the ‘PW, BIS DWDU Notifications’ ([RA-BISDWDUNotify@pa.gov@pa.gov](mailto:RA-BISDWDUNotify@pa.gov@pa.gov)) email address.

## Standards:

### Access Standards

All individuals requiring Informatica access for EDW development must complete and submit an 'Informatica Developer Registration Form' to the EKMS Informatica Team.

1. Only one account will be allowed per individual
2. All individuals must have a CWOPA ID
3. All forms must be filled out correctly and in their entirety as per the instructions on the form - incomplete and/or illegible forms may cause delays

Once received, all forms should be processed by the EKMS Informatica Team within five (5) business days.

### EDW Informatica Standards

#### Mapping:

1. Naming Convention:
  - a. use an underscore in between each differentiation
  - b. m – mapping
  - c. function – such as insert, delete, update, extract, stage
  - d. target table name
  - e. summary detail, purpose of mapping, or column updated

ex. m\_update\_t\_claim\_dim\_MCO\_Plan\_code  
m\_insert\_eligty\_fact\_co\_51\_records
2. Description:
  - a. In the General tab, give a detailed description of what is occurring inside the mapping.
3. Techniques:
  - a. SQL overrides
    - i. Only use when necessary such as to pass a date parameter that works as a filter on the data.
    - ii. Unions will be evaluated and may not be accepted as efficient coding. You should not use unions unless absolutely necessary.
    - iii. No hard coded dates and no changing of the date formats. This should be done later in the transformation.
    - iv. Use an 'ORDER BY' in the SQL so you start with ordered data (if sorted data is needed further on in your transformation). This helps with performance.
    - v. Use when pulling from more than one databases in the 'WHERE IN' clause.
  - b. Expression
    - i. Name it starting with EXP\_
    - ii. Use the Expression transformation to calculate values in a single row before you write to the target
    - iii. To perform any non-aggregate calculations
    - iv. To test conditional statements before you output the results to target tables or other transformations
    - v. To store variable data values
  - c. Filter
    - i. Name it starting with FIL\_
    - ii. Only rows that meet the condition pass through the Filter transformation.

- iii. Filter out unnecessary data when doing updates. Do not update rows with the same value in the current data.
  - d. Sequence Generator
    - i. Name it starting with SEQ\_
  - e. Stored Procedure
    - i. The Stored Procedure transformation name is the same as the stored procedure you selected.
    - ii. Do not create without first consulting with the EKMS Informatica team.
  - f. External procedure
    - i. Should not be used without first consulting with the EKMS Informatica team.
    - ii. The names of the ports, datatypes and port type (input or output) must match the signature of the external procedure.
  - g. Custom transformation
    - i. Name it starting with CT\_
    - ii. Custom transformations operate in conjunction with external procedures. Input and Output ports are handled in separated transformations.
  - h. Normalizer
    - i. Name it starting with NRM\_
    - ii. Used when the sources are COBOL to normalize the 'OCCURS' clause.
  - i. Router
    - i. Name it starting with RTR\_
    - ii. Router transformation tests data for one or more conditions and gives the option to route rows of data that do not meet any of the conditions to a default output group
    - iii. Use in place of filter when you need to pass rows for more than 1 condition or/and when rows need passed that do not meet the condition.
  - j. SQL transformation
    - i. Name it starting with SQL\_
    - ii. Use to run a SQL query where the output will become part of the data flow. Also used when you need to create tables during the mapping run. Make sure the query is efficient. Build the query in a SQL editor and analyze the explain plan.
    - iii. The query has to be one continuous line. It should not have any line breaks except at the very end.
  - k. Joiner
    - i. Name it starting with JNR\_
    - ii. If using a joiner instead of a lookup, make sure to select the smaller table as the master table.
    - iii. Use to join two (2) flat files or two (2) different entity types as sources.
    - iv. Consider the amount and size of flat files being joined. It may be better to stage the flat files first, rather than performing more than one join within a mapping.
    - v. May be used to join two (2) tables together, if the one needs to be grouped by using the aggregator and the other uses plain SQL, then the two (2) tables get joined together by using the joiner.
  - l. Update strategy
    - i. Name it starting with UPD\_
    - ii. Do not use unless it is necessary. If the mapping only updates – Just set the session property to UPDATE and do not use the update strategy.
    - iii. Type II Dimension mappings must use an update strategy.
    - iv. When using an update strategy insert detailed documentation inside the General tab of the mapping and any sessions, worklets and workflows.
    - v. Update strategies hinder performance by taking the power of the database software away.
  - m. Aggregator object
    - i. Name it starting with AGG\_
    - ii. Always use sorted input.

- iii. If performance is an issue, an option would be to give the aggregator its separate cache. This may only be done via written approval by the EKMS Informatica team.
  - n. Rank object
    - i. Do not use as a sorter works better.
    - ii. If must use, Name it starting with RNK\_
  - o. Lookups
    - i. Name it starting with LKP\_
    - ii. Have your data sorted before it comes into the lookup.
    - iii. Does an automatic order by statement on the data.
    - iv. Persistent lookup – for a lookup is going to be used several times inside a workflow. Inside the first mapping that uses the lookup, by checking persistent in the lookup object, this lookup will only be cached once versus re-caching the same lookup for each mapping that is calling the lookup.
  - p. Re-usable object
    - i. You can design a transformation that you can reuse in multiple mappings or mapplets within a folder, a repository, or a repository domain.
    - ii. Non-reusable transformations exist within a single mapping.
    - iii. Reusable transformations can be used in multiple mappings.
    - iv. When you add instances of a reusable transformation to mappings, you must be careful that changes you make to the transformation do not invalidate the mapping or generate unexpected data.
  - q. Mapplets
    - i. Name it starting with MPLT\_
    - ii. Create a mapplet when there is an exact duplicated process that will need to be used for several mappings.
    - iii. Do not create a mapplet that is a combination of multiple lookups or functions that may or may not be used in several mappings. This creates unnecessary processing that must occur each time a mapping runs. For example you are developing ten (10) mappings and each of those mappings uses ten (10) different lookups. Do not create a mapplet that opens all those ten (10) lookups, but only an individual mapping will use only one of the lookups. This creates extra processing and uses unnecessary cache space that likewise could be used for other mappings that are processing at the same time.
- 4. Flat Files and Staging tables
  - a. If a flat file is going to be used by more than one (1) mapping, it is probably better to create the flat file as a staging table especially if the flat file is a large file.
    - i. Staging tables
      - 1. If they are needed, please include the space needed into your capacity plan
      - 2. If you have many staging tables to be created than request your own tablespace through database, when requesting your space allocation. If you do not request your own tablespace, then you are using the stage\_data that is available with the TRANSFORM object. This space is used by all developers and controlled by the EKMS Informatica Team. If you do not properly name the table and request your space, you risk having your table dropped without notice.
      - 3. Name the staging tables starting with the project name. This helps determine who is responsible for the staging table.
      - 4. When creating the staging table, put a description and the name of the person responsible for the table in the comments.
      - 5. If the staging table is for testing, then the name of the table should start with the Informatica Username of the person creating/using the table.  
Ex. PeterG\_t\_claim\_dim\_test
- 5. Team Based Development
  - a. Deployment folder – deployment groups are not being used. To deploy – the whole folder is moved from Development to Production.

- b. The actual folder move of the folder to TFP is to be coordinated by the EKMS Informatica Team. See 'Deployment Standards'.
6. Completion of the 'DW Iconic View and Detail Document' document for each mapping
- a. Iconic view of entire mapping.
  - b. Paragraph detailing the working of the mapping. This gets attached to the iconic view of the mapping. Give enough information that anyone who looks at these two (2) documents together will get a quick grasp on all the sources, targets and what is occurring inside the mapping.
  - c. Include following information:
    - i. Are the data sources flat files
      - 1. Location of the flat files
        - a. from what system and server
      - 2. Name of files
      - 3. Storage of flat files
        - a. what server and what folder – used for loading
        - b. What server and what folder – for archiving, if any? How long are the files archived?
      - 4. What time does the mapping run to use the source files?
      - 5. Are there dependencies that must occur first before these source files are loaded?
      - 6. Are they loaded into staging tables and in what environment?
      - 7. Are the data sources relational database tables
        - a. From what system and what database?
        - b. Name of tables
        - c. Oracle or SQL server
        - d. Database userid and Password
7. Quality Control
- a. This review consists of the following but not limited to:
    - i. Are all Standards followed:
    - ii. Are all techniques used appropriately
    - iii. Performance tests
    - iv. Iconic view with paragraph description of what is occurring during the mapping run attached.
    - v. Above paperwork has all been submitted and questions answered.

### **Tasks:**

- 1. Naming Convention
  - a. use an underscore in between each differentiation
  - b. tsk– task
  - c. include function/action – such as timer, email, command

### **Sessions:**

- 1. Naming Convention
  - a. use an underscore in between each differentiation
  - b. s - session
  - c. include function/action – such as update, insert, delete,

### **Worklet:**

- 1. Naming Convention
  - a. Use underscore between each differentiation
  - b. wklt – worklet
  - c. Function – logical name of grouping of mappings putting together under the worklet. Also, if worklet is performing function such as update, insert, delete, stage, please include in the name.

## **Workflow:**

1. Naming Convention
  - a. Use an underscore in between each differentiation
  - b. wf – workflow
  - c. Function – logical name for grouping underneath workflow. What system, what function.
    - i. Such as
      1. wf\_PROMISE\_monthly\_update\_1
      2. wf\_promise\_monthly\_update\_2
      3. wf\_encounter\_recipient\_update

## **EDW Databases**

Please refer to the 'Data Warehouse Landscape' document for database instance specifics.

## **Deployment Standards**

All Data Warehouse deployments to TFP will be coordinated by the EKMS Informatica Team. The team is to be notified via the 'PW, BIS DWDU Notifications' email address a minimum of ten (10) business days before the move is to occur. The request should be submitted via the 'DW Informatica Developer Special Request - Issues Form'. EKMS will then follow-up with the requestor to coordinate the deployment. All deployments must adhere to Knowledge Management Standards. If it is determined post-implementation that a deployment does not adhere to Knowledge Management Standards, all future deployments pertaining to the project will be placed on hold until such time as the code in question is brought up to standards.

## **Scheduler Standards**

All Data Warehouse mappings, tasks, sessions, mapplets, workflows, scripts, etc. will be fully automated and scheduled via the Informatica Integration Service scheduler that is included with the Informatica PowerCenter suite of tools. No other schedulers or scheduling methods should be used unless prior approval is obtained from EKMS. In the event the use of an alternate scheduler is granted, OpCon will be the only substitute allowed.

## **Exemptions from this Standard:**

There will be no exemptions to this standard.

## **Refresh Schedule:**

All standards and referenced documentation identified in this standard will be subject to review and possible revision annually or upon request by the DHS Information Technology Standards Team.

## Standard Revision Log:

Change Date	Version	Change Description	Author and Organization
7/22/2008	1	New document	BIS Enterprise Knowledge Management Section
2/19/2016	1.1	Reviewed for content. Changed 'Public Welfare' to 'Human Services'; changed 'DPW' to 'DHS'; removed reference to ODS source data; removed details references to database instances, referred users to 'DW Landscape' document, removed the version number from Informatica, provided additional detail under 'Access Standards', removed outdated 'Source Database Standards' paragraph, updated contact email address, and updated the 'Scheduler Standards' paragraph.	Don Pidich – EKMS