README

This README provides information to setup and run the Pentaho Kettle jobs and transformations for the Malawi Central Data Warehouse (CDW).

Things covered:

- General Information
- Installation of Pentaho Kettle Software
- Setup of CDW Code Repository
- Configuring Database Connections
- Running IDS ETL Batch from Command-line
- Modifying ETL Code
- Summary of IDS ETL Jobs and Transformations
- Additional Information

General Information

- Pentaho code: /opt/pentaho/
- Repository configured: /opt/pentaho/CDWRepository
- Server script to run repository code: CDWRun.sh
- TFS Repository: DMTProjects/CDW

    - CDWRepository - Developed Pentaho Code
    - data-integration - Pentaho PDI Software v8.3
    - SQLDeploy - SQL Scripts for DB updates and test data

The /opt/pentaho folder is owned by the user pentaho and is part of the pentaho group. Others needing access to this folder should be added to the pentaho group by an admin. The following is an example of adding a user to the pentaho group:

```
usermod -G pentaho {user}
```

Some the file extensions used by PDI for code files are:

- kdb - Shared Database Connection
- kjb - Job
- ktr - Transformation

Installation of Pentaho PDI Software

Installing PDI

The latest version of the community edition of Pentaho Data Integration (PDI - aka Kettle) software can be found here: https://sourceforge.net/projects/pentaho/files/latest/download

Prerequisites

PDI requires the Oracle Java Runtime Environment (JRE) or Oracle Java DevelopmentKit (JDK) version 8. You can obtain a JRE or JDK for free from Oracle.

Installation

PDI does not require installation. Simply unpack the zip file into a folder of your choice.

On Unix-like operating systems, you may need to make the shell scripts executable by using the chmod command:

```
cd data-integration
chmod +x *.sh
```

Running

PDI comes with a graphical user interface called Spoon, command-line scripts (Kitchen, Pan) to execute transformations and jobs, and other utilities.

The graphical user interface (Spoon) requires a console which supports a graphical interface. To run Spoon on linux it will require an X windows console. Therefore, it is recommended installing software also on a windows desktop to make it easier to run Spoon and maintain ETL custom code. Once code changes are done they can be run via command-line on any designated machine with Pentaho installed.

Note: Pentaho software settings are stored within a directory called ".kettle". The first time Spoon is run this directory will automatically be created if not already there. The location of this directory, by default, is located one directory up from the data-integration sub-folder.

Setup of CDW Code Repository

Pentaho shares files and folders through the Pentaho Repository. The Pentaho Repository is an environment for collaborative ETL development. A one-time setup must be done to connect an environment to the file repository where the developed Pentaho code has been place.

This is primarily done for the development environments where the code will be maintained. Once code changes have been updated, all modified files in the repository can simply be transferred to the location where it will be typically ran from.

To setup / connect to a file repository, do the following steps:

1. Within Spoon, select the "Connect" option located at the top right corner.
2. On the "New Repository Connection" dialog window, select the "Other Repositories" option (NOT Get Started).
3. Select the "File Repository" option and then select "Get Started".
4. For the Repository Details provide the following:

   - Display Name: "CDW" (or other meaningful name to show)
   - Location: (Browse and select the location where the developed code has been placed - for example: C:\DevSource\CDW\CDWRepository)
   - Launch connection on startup: (Check box to true)

5. Select the "Finish" button and then select the "Connect Now" button.

Once the connection to the file repository has been established you can easily access files within the repository by either selecting the "Explore Repository" menu icon (third from the left) or via the menu at Tools->Repository->Explore... or the shortcut key CTRL-E.

Note: Repository definitions are stored within Pentaho's ".kettle" settings folder. The first time Spoon is run this directory will automatically be created if not already there. The location of this directory, by default, is located one directory up from the data-integration sub-folder. Repository definitions are stored within the repositories.xml file.

Configuring Database Connections

Prior to running the Pentaho transformations for the first-time (or after any DB connection changes), the database connection details for Pentaho need to be reviewed and updated if needed.

If using Spoon, this can be done via the Repository Explorer (CRTL-E). Select the Connections tabs and review and edit each of the defined database connections by selecting the connection and clicking the edit icon (pencil).

The connections definitions can also be updated directly by modifying the text files within the repository. They are the files with file extensions ".kdb".

Running IDS ETL Batch from Command-line

The shell script /opt/pentaho/CDWRun.sh can be used to run the IDS ETL job within Linux. The contents of the script is as follows:

```
/opt/pentaho/data-integration/kitchen.sh -rep=CDWRepository -dir=/ -job=main
```

A user should be assigned the pentaho group to run the script or pentaho jobs on the server.

Modifying ETL Code

Pentaho ETL code should be modified with the Spoon UI. To run Spoon on linux a GUI interface (X-Windows) must be configured. It might be easier to have a PDI installation on a Windows computer to modified the ETL code. A copy of the file repository can be kept locally and transferred after changes (or a shared drive could be used if available).

Summary of IDS ETL Jobs and Transformations

- main - Job - This is the primary job to to run all the other jobs and transformations which update the IDS with the latest data from the RDS.
- main.ids - Job - Started by the Job main to run the transformations to update the IDS.
- etl.setParameters - Transformation - Converts values from the table audit_etl_parameter into parameters for Pentaho to use when running the rest of the jobs and transformations.
- etl.writeParametersToLog - Transformation - Writes requested parameters to output log.
- ids.appointment - Transformation - Updates the IDS appointment table.
- ids.breasFeedingStatuses - Transformation - Updates the IDS breastfeeding_statuses table.
- ids.calcLastUpdateDate - Transformation - Calculates the date to use for pulling deltas from the source for all the target tables and creates parameters containing the date for each target transformation.
- ids.contaceDetails - Transformation - Updates the IDS contact_details table.

- ids.diagnosis - Transformation - Updates the IDS diagnosis table.
- ids.encounters - Transformation - Updates the IDS encounters table.
- ids.family_plannings - Transformation - Updates the IDS family_plannings table.
- ids.labOrders - Transformation - Updates the IDS lab_orders table.
- ids.medicationAdherences - Transformation - Updates the IDS medication_adherences table.
- ids.medicationDispensations - Transformation - Updates the IDS medication_dispensations table.
- ids.outcomes - Transformation - Updates the IDS outcomes table.
- ids.patientHistories - Transformation - Updates the IDS patient_histories table.
- ids.people - Transformation - Updates the IDS people table.
- ids.personAddresses - Transformation - Updates the IDS person_addresses table.
- ids.personHasTypes - Transformation - Updates the IDS person_has_types table.
- ids.personNames - Transformation - Updates the IDS person_names table.
- ids.pregnantStatuses - Transformation - Updates the IDS pregnant_statuses table.
- ids.presentingComplaints - Transformation - Updates the IDS presenting_complaints table.
- ids.relationships - Transformation - Updates the IDS relationships table.
- ids.sideEffects - Transformation - Updates the IDS side_effects table.

Additional Information

- Pentaho Documentation: https://help.pentaho.com/Documentation
- Pentaho CE Site: https://sourceforge.net/projects/pentaho/