Code Page Settings and Performance Settings for the Data Validation Option
Abstract
This article provides general information about code page settings and performance settings for the Data Validation Option. Configure these settings to avoid translation errors during archive and validation of data.

Supported Versions
- Data Validation Option 3.x

Table of Contents
Introduction .......................................................................................................................... 2
Code Page Settings .......................................................................................................... 2
  PowerCenter Services .................................................................................................... 2
  PowerCenter Client Code Pages .................................................................................... 5
  Data Integration Service Machine Configuration ......................................................... 6
  ILM Web Console .......................................................................................................... 7
Configuring Data Validation Option and PowerCenter Performance Settings ............. 8
  Data Validation Option Settings .................................................................................. 8
  PowerCenter Settings ................................................................................................. 8
Appendix A. Modifying a Repository Code Page ............................................................. 10
Appendix B. Determining When to Split Data Validation Option Jobs into Parts .......... 11
Appendix C. Calculating Disk Cache Requirements for a Data Validation Option Session... 12
  Steps to Use the Spreadsheet ..................................................................................... 12

Introduction
This article provides general information about code page settings and performance settings for the Data Validation Option. You might need to configure settings in Data Archive and PowerCenter. Configure these settings to avoid translation errors during archive and validation of data.

Code Page Settings
You set data movement modes and code pages for PowerCenter services, in PowerCenter client tools, and on the Data Integration Service machine.

PowerCenter Services
You set data movement modes and code pages for the PowerCenter Integration Service and the PowerCenter Repository Service.
Note: You do not set environment variables for LANG and NLS_LANG in the properties of the PowerCenter Integration Service or PowerCenter Repository Service. You set them on the machines hosting the PowerCenter Integration Service and PowerCenter Repository Service.

PowerCenter Repository Service Code Page
When creating a new Repository Service, you will be prompted to select the code page.
The following figure shows an example of selecting a code page during PowerCenter Repository Service configuration:

![Create New Repository Service](image)

For an Application Archiving project, set the code page to UTF8 to ensure better character conversion.

After you create the Repository Service, you cannot change the repository code page. If you need to modify the code page, then you must back up the repository, delete the Repository Service, create a Repository Service with the same name and database credentials, and restore the repository contents from the backup. For more information, see Appendix A. Modifying a Repository Code Page.
PowerCenter Integration Service Code Page

You set the data movement mode for the Integration Service to Unicode in the General Properties Settings on the Properties tab:

Set the code page for the Integration Service node to UTF-8 on the Processes tab:
**PowerCenter Client Code Pages**

**Workflow Manager**

In the Workflow Manager, click Connections > Relational to set the code page of the relational database connection to UTF-8:

The Code Page option appears in the connection properties:

![Connection Object Definition](image)

The following table shows a summary of the typical settings:

<table>
<thead>
<tr>
<th>Source</th>
<th>Source Code Page</th>
<th>PowerCenter Source Relational Connection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>Various (ISO 8859-1, UTF-8, WEBMSWIN1252)</td>
<td>UTF8</td>
</tr>
<tr>
<td>DB2 Z/O S</td>
<td>Not confirmed</td>
<td>UTF8</td>
</tr>
<tr>
<td>DB2 AS/400</td>
<td>Not confirmed</td>
<td>Not confirmed</td>
</tr>
<tr>
<td>MS SQL</td>
<td>UTF16</td>
<td>UTF16</td>
</tr>
<tr>
<td>JDE World</td>
<td>Various</td>
<td>UTF8</td>
</tr>
<tr>
<td>FAS (odbc)</td>
<td>UTF8</td>
<td>UTF8</td>
</tr>
</tbody>
</table>
Data Integration Service Machine Configuration

On the machine hosting the Data Integration Service, you need to set the code page in several places. The following information is based on Red Hat Linux and may be slightly different for your operation system:

.profile Settings

The LANG and LOCAL settings in the .profile settings must match the code page entry on the server. Enter the following command to verify the setting:

    locale -a

Find the entry similar to en_US.utf8 and use this value in the .profile settings. For example, this might be en_US.utf8 or en_US.UTF8.

In the .profile (or similar file) under the directory of the PowerCenter owner user ID, add the following entries:

    NLS_LANG=AMERICAN_AMERICA.AL32UTF8
    LANG=<code page entry>
    LOCAL=<code page entry>
    DB2CODEPAGE=1208

For example:

    NLS_LANG=AMERICAN_AMERICA.AL32UTF8
    LANG=en_US.utf8
    LOCAL=en_US.utf8
    DB2CODEPAGE=1208

Note on the NLS_LANG Setting

Querying the Oracle database using the following query returns information of interest for setting the NLS_LANG value:

    select * from V$NLS_PARAMETERS

Values for NLS_LANGUAGE, NLS_TERRITORY, and NLS_CHARACTERSET are used in the components. The resulting NLS_LANG value needs to match the PowerCenter Repository database values.

odbc.ini

In the odbc.ini file, set the following:

FAS Connections:

Add the following IANAAppCodePage=106

Also append the following to the CustomProperties:

|--no-disable-wide-chars

For example:

    [ABCD_201011]
    Driver=/app/prod/ilm/File_Archive/npa_home/ODBC/RainStor-ODBC/lib64/ivoa22.so
    Description=RainStor 2.0 ODBC driver
    Host=123.456.78.90
    Port=3731
    ServerDataSource=
    UseLDAP=0
    DistinguishedName=
    Encrypted=0
    LoadBalancing=0
AlternateServers=
ConnectionRetryCount=0
ConnectionRetryDelay=3
CustomProperties=Archive=ABCDRS_201011;Parser=SQL92|--no-disable-wide-chars
DescribeWcharAsChar=1
DescribeVarcharAsLongvarchar=4096
IANAAppCodePage=106

**JD Edwards Connections**

Add the following entries to DE Connection entries:

- CharsetFor65535=**2028**
- IANAAppCodePage=**2028**

The value to the right of the equal sign on the parameters above must match the JDE code page. Set to whatever matches the particular JDE Connection data. The value in the example above, 2028, may vary depending on the source data.

Append the following to the CustomProperties:

- |--no-disable-wide-chars

Archive them with IBM037 (IANA 2028) from iWay.

The CharsetFor65535 and the code page in iWay need to correspond and may be different for connections to different JDE data. Below is the IANAAppCodePage values provided by Datadirect: [http://www.datadirect.com/support/troubleshooting/su-faq-iana.html](http://www.datadirect.com/support/troubleshooting/su-faq-iana.html)

**JDBC Driver**

Use the Oracle 11g JDBC driver. We have seen issues with the 10g driver.

**ILM Web Console**

Add the following for the source connections:

```
IWASERV;CFG=/app/dev/ilm/jesamp02.cfg;
```

![ILM Web Console](http://www.informatica.com)

**New/Edit Archive Source**

- **Connection Name**: JDE_QUADRO_SRC
- **Description**: JDE_QUADRO_SRC
  - **ILM Service Host**: gr190032.gr-assm.ymail
  - **ILM Repository Path**: /tmp
  - **Server**: user1🔗@https://gr190032.gr-assm.ymail
  - **ILM Repository Administrator User**: ADMIN
  - **Admin Login Name**: ADMIN
  - **Password**: ********
  - **Calendar Password**: ********
  - **Application User Name**: ADMIN
Configuring Data Validation Option and PowerCenter Performance Settings

Several settings were required in order to optimize the performance of large Data Validator PowerCenter jobs. Some of the settings involve configuration for Data Validation Option and some are settings in PowerCenter.

Data Validation Option Settings

.profile Settings

On the server where Data Validation Option is installed, add the following environment variable to the .profile or similar file for the user ID that installed Data Validation Option:

```
DV_TRANS_CACHE_SIZE_AUTO = y
```

You must restart Data Validation Option in order for the value to be effective. This value will set all sorter, aggregator, joiner and other cache sizes to Auto for the Index and Data cache size.

Data Validation Option Preferences

In the Data Validation Option Utility, set the Mapping Preferences to set the DTM Buffer Block Size to Auto:

PowerCenter Settings

Session Settings for Large Tables in the Workflow Manager

Some of the settings in a PowerCenter session will need to be modified for better performance. Data Validation Option creates a session configuration for each Data Validation Option session and creates the session to point to it instead of the default_session_config for the PowerCenter folder. Override this value in the session for jobs that are large or that are performing poorly. Defining a large job is not straightforward because it includes a combination of many rows, many columns, or columns that are large [e.g. VARCHAR2(4000)].

Note: The settings specified were determined on a Linux server with 64 GB of RAM and 600 GB or more of disk space available to the cache directory.
Set the DTM Buffer Size property on the Properties tab of the session properties:

Although we have suggested that in general the DTM Buffer Size be set to Auto, for large jobs, set it to 2 GB, defining it using all numeric values (2000000000).

On the Config Object tab, set the following recommended values for the following properties:

- Default Buffer Block Size property to 3MB
- Maximum Memory Allowed for Auto Memory Attributes to 2GB
- Maximum Percentage of Total Memory for Auto Memory Attributes to 10.

Adjust these as needed and depending on the amount of memory available on the server.
On the Mapping tab, select the Joiner transformation, set data cache size to 40GB and index cache size to 5GB. If there is more than one Joiner transformation, adjust to the value to share server memory accordingly:

Appendix A. Modifying a Repository Code Page

**PowerCenter Repository Service**

**Note:** After a repository is converted to Unicode (UTF-8) code page it cannot be converted to any other character set (ISO-8859-1, etc.).

1. Login to the Administration Console.
2. Disable all Integration Services associated with this Repository Service.
3. Ensure that no user connections are visible under the Connections tab.
4. Backup the repository using the Administration Console > Repository > Backup Contents (enter the Administrator username/password and the backup filename).
5. Change the Operating Mode to Exclusive.
6. Delete the contents of the repository.

**Note:** You can also create a repository in another schema and test if the backup file can be restored correctly.

1. Change the Repository Service Operating Mode to Normal and disable the repository.
2. For PowerCenter 8.1.x, change the repository code page under Repository > Database Properties > Edit > Codepage.
   For PowerCenter 8.6.x and 9.0.x, delete the Repository Service and recreate it with the same name but with the correct codepage of Unicode (UTF-8).
3. Shut down the PowerCenter domain.
4. Change the operating mode, code page and shut down the PowerCenter domain of all nodes in the domain.
5. **(Oracle)** Set the NLS_LANG environment variable to `AMERICAN_AMERICA.UTF8` or `AMERICAN_AMERICA.AL32UTF8` (depending on the database character set).

6. **(UNIX)** Set the `LC_ALL` environment variable to `en_US.UTF-8`.

7. Restart the node using `infaservice.sh` startup.

8. Enable the repository (a message saying *Repository has no contents* should be displayed).

9. Restore the repository contents from the backup file (use the same Administrator user name and password used to create the backup file).

**Note:** After the code page conversion, any migration from this repository can only be to another Unicode repository (either using folder copy or using object export).

**PowerCenter Integration Service**

If the Codepage for the Integration Service also needs to be changed, then you can perform the following steps additionally.

1. Change the `DataMovementMode` mode to "Unicode" in the Integration Service.
2. Click the `Processes` tab of the Integration Service.
3. Change the codepage to `UTF-8 encoding of Unicode` for all nodes that are assigned to this Integration Service.

**Appendix B. Determining When to Split Data Validation Option Jobs into Parts**

For very large tables, cache resources available on the server may not be adequate to perform validation successfully. In this case, the processing for one table may be split into parts, each processing only a portion of the columns for the table. Cache resources include a combination of memory and disk space. Some PowerCenter transformations create cache, including Sorter and Joiner transformations, which might be part of the DVO-generated mapping.

To determine when the table should be processed in parts, first determine the overall size of the table, as seen in PowerCenter. PowerCenter works with VARCHAR as a string, and thus holds space for very large VARCHAR2 fields, even if they are not utilized. To begin, in the PowerCenter Designer, open the mapping or the source in PowerCenter (Source Analyzer or Mapping Designer) and open the source so that all ports are visible:
Highlight the entire list of ports by clicking in the column number field and holding as you scroll through the list, or click in the port number 1 field, and then Shift-Click in the last port number:

Press CTRL-C to copy and paste into an MS Excel worksheet. Sum the width of all columns. Multiply this by the number of total rows in the table. If this value exceeds the amount of memory + disk space available on the server, or leaves no room for other transformations or memory required for the job (e.g. DTM Buffer) then split the job into parts. Often separating the large VARCHAR2 fields is adequate to run the job successfully.

Appendix C. Calculating Disk Cache Requirements for a Data Validation Option Session

The attached spreadsheet calculates the amount of disk cache required for Data Validation Option sessions in PowerCenter. This provides a general idea for the settings, not an exact value. Perform for all sessions that run concurrently.

Steps to Use the Spreadsheet

1. Run a session with 100 rows. Get the session log.
2. Open the spreadsheet; go to the second sheet "Sample".
3. Fill in cell F1, G1, H1 using the information in the session log.
4. Get an estimates of the source row count, e.g. do a “select count(*)” on the source table. Fill in the number in column E.
5. Go to sheet 1, column B shows you the disk cache size estimations and the number of cache files.

Example
Cache size from one of the joiners as an example.

Step 3
DBG_21694 : NAME_5 Partition [0]: Index cache size = [715827882], Data cache size = [1431655764]

Take the higher number of the 3 messages above and put it in column F in sheet 2. In this case, index cache size is 715831600, and data cache size is 1431671808.

CMN_1795 : The index cache size that would hold [12161420] input rows from the master for [NAME_5], in memory, is [1092036000] bytes.

Take the first number and put it in column G in sheet 2. In this case, the number of input rows is 12161420.

CMN_1794 : The data cache size that would hold [12161420] input rows from the master for [NAME_5], in memory, is [261519175680] bytes.

Take the first number and put it in column H in sheet 2. In this case, the memory need is 261519175680.

Step 4
Let's say the source table has 27 million rows. Put this number in column E in sheet 2.

Finally
Go to sheet 1 in the attached spreadsheet to see the minimum disk space required for this session. Do this for all the joiners and aggregators. In the customer's case, the joiner is using a lot of disk space and the session failed because other sessions are creating many large disk cache files as well.

Author
Janet Sitchin
IPS Senior Consultant