ETL Systems; XML Processing in PHP

May 11, 2013

1 ETL - principles, applications, tools

1.1 ETL: Extract-Transform-Load

Extract-Transform-Load (ETL) are data integration practices and tools:

Extract data-mining from different sources, different data formats, ...

Transform transformation of data to a desired form

Load loading/storing of data to/from a target database/data warehouse

1.2 ETL Applications

ETL tools has many application areas today:

1. Different sources and formats data integration (text documents, CSV, XLS spreadsheets, databases, XML data, ...)
2. Data consolidation (transformations of data and data "cleaning")
3. Storing of data into huge databases - data warehouses for management applications
4. Data migration (data transfers to different platforms, databases, etc.

ETL systems are called as "a critical building block to a successful business intelligence deployment".

1.3 Implementation

There is a lot of (not only java-based) implementations, many have a GUI, allowing to graphically design transformation flows.

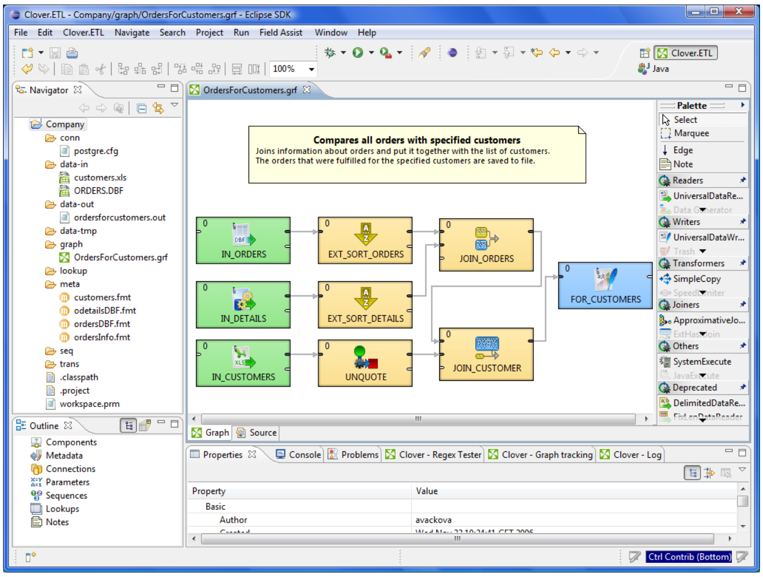
Clover ETL `http://www.cloveretl.org` - open source ETL tool including GUI (`http://www.cloveretl.org/\_img/clovergui/Graf.png`)

Microsoft SQL Server Integration Services `http://www.microsoft.com/sql/technologies/integration/default.mspx`

Octopus Java/XML ETL Tool `http://octopus.enhydra.org/`

java-etl `http://code.google.com/p/java-etl/`

Kettle `http://kettle.pentaho.org/`



Compares all orders with specified customers
Joins information about orders and put it together with the list of customers.
The orders that were fulfilled for the specified customers are saved to file.

Property	Value
Basic	
Author	avackova

1.4 Examples - Clover ETL

Commercially developed (Javlin (`http://www.javlin.cz`) company, FI industry partner, `http://www.cloveretl.org`) open-source tool containing:

Clover Engine The kernel processing transformations. Contains connectors to external data sources and targets.

Clover Server deployment platform for transformation execution (incl. planing and monitoring) in a real life.

Clover Designer tool for graphical design of transformation graphs (based on Eclipse platform).

1.5 Clover ETL - Designer

1.6 Questions

ETL implementation and deployment on a huge data involves some problems, that are not usual in a different areas:

- transformations should be optimized to a speed as well as to allow huge data processing.
- effective memory models used to (temporal) store XML data - unable to use common "in memory" tree models.
- definability, maintainability and verifiability of wide transformation networks - visual tools + formal methods

2 XML interfaces for PHP

2.1 Concepts

Principially the same as in Java, there are:

tree-oriented interfaces DOM (`http://php.net/manual/en/book.dom.php`) full repertoire of operations (read, validate, write incl. prettyprinting, programmatic creation of docs, elements, etc.)

stream-based (pull) SimpleXML (`http://php.net/manual/en/book.simplexml.php`) - since PHP 5.0 part of the core PHP, very simple and frequently used interface, enables direct iteration (traversal) through XML elements, direct evaluation of XPath expressions etc.Also see SimpleXML (PHP) at W3Schools (`http://www.w3schools.com/php/php\_xml\_simplexml.asp`)

event-driven SAX (`http://php.net/manual/en/book.xml.php`) - similarly as in Java, in all recent PHP compilations

2.2 Example (1) - DOM

The following code reads (analyses, "parses") XML document and writes it back to file (serializes it).

```
$dom = new DOMDocument();

// configuration for read
$dom->preserveWhiteSpace = FALSE;
$dom->load('input.xml');

// configuration for write
$dom->formatOutput = TRUE;
$dom->encoding = 'utf-8';
$dom$\to$save('output.xml');
```

2.3 Example (2) - SAX

The following code reads an XML file with book records and prints info about htem them (from Reading and writing the XML DOM with PHP Using the DOM library, SAX parser and regular expressions, Jack Herrington, IBM 2005)

```
<?php
  $g_books = array();
  $g_elem = null;

  function startElement( $parser, $name, $attrs )
  {
  global $g_books, $g_elem;
  if ( $name == 'BOOK' ) $g_books []= array();
  $g_elem = $name;
  }
  function endElement( $parser, $name )
  {
  global $g_elem;
  $g_elem = null;
  }

  function textData( $parser, $text )
  {
  global $g_books, $g_elem;
  if ( $g_elem == 'AUTHOR' ||
  $g_elem == 'PUBLISHER' ||
  $g_elem == 'TITLE' )
  {
  $g_books[ count( $g_books ) - 1 ][ $g_elem ] = $text;
  }
  }

  $parser = xml_parser_create();

  xml_set_element_handler( $parser, "startElement", "endElement" );
  xml_set_character_data_handler( $parser, "textData" );

  $f = fopen( 'books.xml', 'r' );

  while( $data = fread( $f, 4096 ) )
  {
  xml_parse( $parser, $data );
  }

  xml_parser_free( $parser );

  foreach( $g_books as $book )
  {
  echo $book['TITLE']." - ".$book['AUTHOR']." - ";
  echo $book['PUBLISHER']."\n";
  }
?>
```

2.4 Example (3) - SimpleXML

From *SimpleXML processing with PHP A markup-specific library for XML processing in PHP by Elliotte Rusty Harold, IBM Developerworks, 2006*

```
<html xml:lang="en" lang="en">
<head>
  <title>XPath Example</title>
</head>
<body>

<?php
```

```
$rss =  simplexml_load_file('http://partners.userland.com/nytRss/nytHomepage.xml');
foreach ($rss->xpath('//title') as $title) {
  echo "<h2>" . $title . "</h2>";
}
?>

</body>
</html>
```

2.5 More (web) resources

DOM Very good (English written) intro to XML in PHP at IBM Developerworks: Reading and writing the XML DOM with PHP (`http://www.ibm.com/developerworks/library/os-xmldomphp/`)

SimpleXML Elliotte Rusty Harold: SimpleXML processing with PHP A markup-specific library for XML processing in PHP (`http://www.ibm.com/developerworks/library/x-simplexml.html`)

XML in PHP Jiří Kosek /in Czech/: Very good Czech-written intro series on XML in PHP Processing by Jirka Kosek (www.zdrojak.cz) (`http://www.zdrojak.cz/serialy/prehled-podpory-xml-v-php5/`)

2.6 More (books)

Jiří Kosek: PHP a XML (in Czech) Grada Publishing, 2010 - excellent, well readable, educative, includes not only info on PHP processing but in general on XML: Schema, Relax NG, XSLT, web services