**Pre-Processing**

Pre –processing of the invoice format file is done in JAVA using NetBeans IDE 8.0.2.

There are multiple tools present in the market for parsing and processing XML files and store the data. Some of them include Simple API for XML (SAX) parser or Document Object Model (DOM) parser or Java Architecture of XML Binding (JAXB). All these parsers require the structures of the XML to be static. Dynamic changes to the tag names or structure cannot be accommodated. The main purpose of our project was to accommodate dynamic structures of an XML file and any changes to the tag names should be accommodated. So usage of any one particular parser was not suiting the requirements.

We have used Node of SAX parser to retrieve the information about XML tag names, data and hierarchy level. This information is stored so that it can be used further. The tags present at level 2 form the tables in the database and their child nodes are the column names of those tables. As the names of the tags are not fixed i.e. they can change with every format, we need to first find out tag name matches which table name. To find that, a connection to the database is made using JBDC and stored procedure is called. This stored procedure will return the table name in which the data should be inserted. Table names for all the tags present at level 2 hierarchy are stored.

Next step is to perform pattern matching for the column names for each table. One column value is inserted at a time. So in order to maintain the key constraints the primary key is first inserted in the table. We have used a status flag in order to find out whether data needs to be inserted in a new row or is a part of the previous row. The flag is set to “new” every time a new row is encountered and is then set to “old”. If the status flag value is “new” then an insert query fired with the table name, column name and data. If the status flag value is “old” then an update query is fired with the same parameters to update the same row but different column values.

In this way the invoice document is parsed and stored in appropriate tables.

**References**

1. <http://ap.finance.ucla.edu/UC_XML_Invoice>.
2. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3920815/>
3. <https://dev.mysql.com/>
4. <https://generateData.com/>
5. <https://www.mysqltutorial.org/>
6. <http://www.databasejournal.com/features/mysql/mysql-fuzzy-text-searching-using-the-soundex-function.html>

**Books**

My SQL Documentation

Modern Database Management

Presentation slides by Prof. Mutsalklisana