**ETL Task**

**Task**: Design an ETL (Extract, Transform, Load) pipeline with met Talend Open Studio for Data Integration (https://www.talend.com/) or a preferred scripting language like R, Python etc. , that imports LOINC files (https://loinc.org/downloads/loinc > LOINC Table File – CSV Text Format ) to a postgresql (http://www.postgresql.org/ ) database according to the i2b2 ontology table format (see https://www.i2b2.org/software/files/PDF/current/Ontology\_Design.pdf ).

**Input**: The loinc.csv and LOINC\_\*\*\*\_MULTI-AXIAL\_HIERARCHY.CSV files.

**Output**: A postgresql table according to the i2b2 ontology table format with a record for each existing LOINC code (including LP prefix codes). The empty table can be created in postgresql, by executing following SQL query:

CREATE TABLE I2B2

( C\_HLEVEL INT NOT NULL,

C\_FULLNAME VARCHAR(700) NOT NULL,

C\_NAME VARCHAR(2000) NOT NULL,

C\_SYNONYM\_CD CHAR(1) NOT NULL,

C\_VISUALATTRIBUTES CHAR(3) NOT NULL,

C\_TOTALNUM INT NULL,

C\_BASECODE VARCHAR(50) NULL,

C\_METADATAXML TEXT NULL,

C\_FACTTABLECOLUMN VARCHAR(50) NOT NULL,

C\_TABLENAME VARCHAR(50) NOT NULL,

C\_COLUMNNAME VARCHAR(50) NOT NULL,

C\_COLUMNDATATYPE VARCHAR(50) NOT NULL,

C\_OPERATOR VARCHAR(10) NOT NULL,

C\_DIMCODE VARCHAR(700) NOT NULL,

C\_COMMENT TEXT NULL,

C\_TOOLTIP VARCHAR(900) NULL,

M\_APPLIED\_PATH VARCHAR(700) NOT NULL,

UPDATE\_DATE timestamp NOT NULL,

DOWNLOAD\_DATE timestamp NULL,

IMPORT\_DATE timestamp NULL,

SOURCESYSTEM\_CD VARCHAR(50) NULL,

VALUETYPE\_CD VARCHAR(50) NULL,

M\_EXCLUSION\_CD VARCHAR(25) NULL,

C\_PATH VARCHAR(700) NULL,

C\_SYMBOL VARCHAR(50) NULL

) ;

**To clarify an example for LOINC code 4548-4 follows**:

• C\_HLEVEL = 9 (code level in the LOINC hierarchy (length from the path from PATH\_TO\_ROOT to LOINC\_254\_MULTI-AXIAL\_HIERARCHY) + 1 (because all entries start with a fixed ‘\i2b2\Laboratory’)

• C\_FULLNAME = The full hierarchy starting from “\i2b2\Laboratory” and further similar to the PATH\_TO\_ROOT column in the LOINC\_254\_MULTI-AXIAL\_HIERARCHY.CSV, Each entry in this hierarchy starts with the ‘\’ character followed by the code between brackets, followed by a space and finally the name of the element. For example for LOINC code 4548-4: \i2b2\Laboratory\(LP31756-7) Hematology\(LP30786-5) Hematology\(LP31617-1) Hemoglobin normal variant\(LP14554-7) Hemoglobin A\(LP16412-6) Hemoglobin A1\(LP16413-4) Hemoglobin A1c\(LP100945-7) Hemoglobin A1c | Bld-Ser-Plas\(4548-4) Hgb A1c MFr Bld

• C\_NAME = "Hemoglobin A1c/Hemoglobin.total:MFr:Pt:Bld:Qn:" (for LP codes this contains the content of the CODE\_TEXT field in LOINC\_254\_MULTI-AXIAL\_HIERARCHY.CSV, for other codes this contains the content from following columns in loinc.csv separated by a ‘:’ : "COMPONENT","PROPERTY","TIME\_ASPCT","SYSTEM","SCALE\_TYP","METHOD\_TYP")

• C\_SYNONYM\_CD = N (fixed)

• C\_VISUALATTRIBUTES = LA (folow the rules defined in https://www.i2b2.org/software/files/PDF/current/Ontology\_Design.pdf)

• C\_TOTALNUM = NULL (fixed)

• C\_BASECODE = ‘LOINC:4548-4’ (prefix “LOINC:” follow by the content of the CODE column in LOINC\_254\_MULTI-AXIAL\_HIERARCHY.CSV )

• C\_METADATAXML = for codes starting with ‘LP’ the content from the similar columns in the LOINC\_254\_MULTI-AXIAL\_HIERARCHY.CSV file and for the other codes the content from the similar columns in the loinc.csv file. This content needs to be transformed in XML by setting as root element the name of the source file (without the extension) and the column name as XML tag and column value as XML value for this tag: <loinc><LOINC\_NUM>4548-4</LOINC\_NUM><!— tags for other columns--></loinc>

• C\_FACTTABLECOLUMN = ‘CONCEPT\_CD’ (fixed)

• C\_TABLENAME = ‘CONCEPT\_DIMENSION’ (fixed)

• C\_COLUMNNAME = ‘CONCEPT\_PATH’ (fixed)

• C\_COLUMNDATATYPE = ‘N’ (‘N’ for LOINC codes with column SCALE\_TYP ‘Qn’, else ‘T’)

• C\_OPERATOR = ‘LIKE’ (fixed)

• C\_DIMCODE = same value as C\_FULLNAME

• C\_COMMENT = NULL

• C\_TOOLTIP = same value as C\_FULLNAME

• UPDATE\_DATE = Date of last record modification (INSERT of UPDATE).

• DOWNLOAD\_DATE = Date when the LOINC source files were downloaded (INSERT of UPDATE)

• IMPORT\_DATE = Date when first records was inserted in the database (INSERT)

• SOURCESYSTEM\_CD = ‘LOINC’ (fixed)

• VALUETYPE\_CD = ‘LAB’ (fixed)

• M\_APPLIED\_PATH = ‘@’ (fixed)

• M\_EXCLUSION\_CD = NULL (fixed)

• C\_PATH = same as C\_FULLNAME without the part following after the last ‘\’ (instructions can be found here https://www.i2b2.org/software/files/PDF/current/Ontology\_Design.pdf)

• C\_SYMBOL = ‘Hgb A1c MFr Bld’ (CODE\_TEXT column in the LOINC\_254\_MULTI-AXIAL\_HIERARCHY.CSV file)