CSC 457 / 557

Project (Part 3)

Points: 50

**PART-A (DUE DATE: 3/28/2020; before 10:00 PM)**

* Develop a high level data staging plan that identifies the sources, targets and data transformations. Turn-in a diagram representing the staging plan along with a document describing the individual transformations.
  + Required data sources
    - SQL-Server
    - MySQL
    - Flat file
    - Spreadsheet
  + Required transformations
    - Merging of information (undergraduate / graduate)
    - Splitting of single fields (undergraduate / graduate)
    - Missing data (filter rows) (undergraduate)
    - Filtering columns (undergraduate / graduate)
    - Decoding of fields (undergraduate / graduate)
    - Derivation (graduate)
    - Missing data (replace with meaningful value) (graduate)
    - Eliminate duplicate records (fuzzy matching) (graduate)

**PART-B (DUE DATE: 4/4/2020; before 10:00 PM)**

* + Create and populate data sources

**PART-C (DUE DATE: 4/19/2020; before 10:00 PM)**

* + Implement your data staging plan using Talend data integration software.
* Requirements:
  + Insert surrogate key values into each dimension table
  + Replace “operational” key values contained in incoming fact table records with CORRESPONDING surrogate key values