EE/CprE/SE 492 – sdmay18-27

Data Analytic Tools for Inconsistency Detection in Large Data Sets

Week 3 Report

02/10/2018 – 02/23/2018

Client/Advisor: Kingland Systems, Ying Cai

Team Members/Role:

Logan Heitz – Project Lead

Christopher Konopka – Communication Lead

Camden Voigt – Technical Lead

Timothy Rogers – Quality Lead

Weekly Summary

Past week accomplishments

- Logan:
    - Created table for storing inconsistencies once they have been detected and implemented methods in the storage module to support the new table
    - Updated record storage so that each reporter has their own table for records they report. Updated the storage module to work with the new table structure
    - Introduced Tuple object that makes incoming records easier to understand in the code.
    - Updated the inconsistency checker to work with the new database design.
    - Moved insertion into the database before the inconsistency check
    - Started implementation of multi-threading for the raw data parse module, so that whenever a new tuple is created a new thread is

started to check inconsistencies on that tuple and the raw data parser will continue to parse records from the XML file.

- Christopher:
    - Implemented check to see if a record already exists in the database before checking inconsistencies. This will avoid doing duplicate inconsistency checks and getting duplicate records into our database.
    - Validate that values in tuple are not empty before checking inconsistencies. If a field in an inconsistency has an empty value that inconsistency should not be checked, as empty fields are a different type of inconsistency.
    - Started work on getting command line options for the program set up to provide another level of configuration for Kingland in starting the program.
- Camden:
    - Begin development of REST API that will be used by the mock UI to display inconsistency information
    - Begin development of UI using HighCharts to display inconsistency information.
- Timothy:
    - Updated the tuple object to be able to save multiple names that is needed for joint account record types
    - Started implementation of multi-threading for the storage module so that calls to the different reporter tables will be done on different thread.
    - Started implementation of multi-threading for the inconsistency checker module so that each of the inconsistency checks will start a new thread.

Pending issues

- Logan:
    - Finish implementation of multi-threading for the raw data parse module, so that whenever a new tuple is created a new thread is started to check inconsistencies on that tuple and the raw data parser will continue to parse records from the XML file.
    - Fix memory leak issue with jdbc statement

- Christopher:
    - Finish command line arguments
    - Update existing uses of tuple to use all strings in the array list instead of just the first string.
- Camden:
    - Finish REST API
    - Finish Front end using HighCharts
- Timothy:
    - Finish implementation of multi-threading for the storage module so that calls to the different reporter tables will be done on different thread.
    - Finish implementation of multi-threading for the inconsistency checker module so that each of the inconsistency checks will start a new thread.
    - Update unit tests to remove SQL dependencies
    - Update unit tests to better work with multi-threading

Individual contributions

Team Member	Weekly Hours	Total Hours
Logan Heitz	16	94
Christopher Konopka	12	78.5
Camden Voigt	14	84
Timothy Rogers	14	87

Plan for coming week

- Logan:
    - Finish implementation of multi-threading for the raw data parse module, so that whenever a new tuple is created a new thread is started to check inconsistencies on that tuple and the raw data parser will continue to parse records from the XML file.
    - Fix memory leak issue with jdbc statement

- Christopher:
    - Finish command line arguments
    - Update existing uses of tuple to use all strings in the array list instead of just the first string.
- Camden:
    - Finish REST API
    - Finish Front end using HighCharts
- Timothy:
    - Finish implementation of multi-threading for the storage module so that calls to the different reporter tables will be done on different thread.
    - Finish implementation of multi-threading for the inconsistency checker module so that each of the inconsistency checks will start a new thread.
    - Update unit tests to remove SQL dependencies
    - Update unit tests to better work with multi-threading

Summary of weekly advisor meeting

02/16/2018 Advisor Meeting

- Checking inconsistencies that have empty values takes time and so we should validate that the values are not empty before checking inconsistencies
- The logger outputting to the console is slowing down our runtime somewhat so we might want to cut down on log statements or have the logger go to a file instead.
- The large example files that Kingland has provided us do not have the namespace in the XML attributes, but the smaller example files do have these namespaces. We'll remove the name spaces from the smaller files for now until we get an XML schema from Kingland.
- We want to switch to a fixed thread pool since SQL only has a set number of threads. We will want to do some tests to figure out what the best number of threads will be.
- Exporting the records to the database before we check inconsistencies will fix any issues with getting inter-report inconsistencies checked and make getting the id of incoming records easier.
- We want to introduce some command line options to further configure the program so Kingland can easily start it.

02/23/2018 Advisor Meeting

- We have credit now to setup an AWS machine so Camden will figure out the configuration that we want and we will get that setup
- Dr. Cai has sent a message to Kingland about getting the XML schema but has yet to hear back from them.
- Memory leak issue with an unclosed jdbc statement that needs to be fixed.