EE/CprE/SE 492 – sdmay18-27

Data Analytic Tools for Inconsistency Detection in Large Data Sets

Week 1 Report

01/12/2018 – 01/26/2018

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:
    - Refactored the existing Export module into new Storage module. This module will then handle all interactions with our storage medium and currently is setup to handle an SQL database. The implementation is abstract and can be extended to other storage mediums for testing without changing other modules in the project.
    - Refactored Export object to into the StorageConfig object that will allow for the setup of the Storage medium based on the settings in the configuration file for the project. Added in the ability to specify the table name that should be used for a database in the configuration file.
    - Added in the QueryBuilder module so that we can easily test the performance of different database queries without making extensive changes to the rest of the code.

    - Created test cases for both the Storage module and the QueryBuilder.
- Christopher:
    - Updated the Inconsistency Matcher to work on inconsistency types that have more than 1 field in the index or compare keys. For example, an inconsistency could be having the same Name and Address as the index, but a different SSN as the compare key.
    - Optimized the query construction for inconsistencies that needed both And and Or for indexing or compare keys. Previously, inconsistencies of the type same SSN or TAX_ID but different Name needed to be done as two different inconsistencies one for SSN and one for TAX_ID. This however requires two queries of the database which is currently a performance bottleneck. Christopher set this up to be a single inconsistency that can then be down in a single query.
- Camden:
    - Created a function for creating a table separate from creating a connection to the database so that we can work with an existing database and not create a new one each time (this was fine for our prototype, but needed to be updated for the full version)
    - Updated existing unit testing functions to provide better code coverage and to get all tests passing for continuous integration
    - Setup continuous integration on Gitlab using the test machine that was provided by our advisor. This can be used to setup automated performance testing in the future.
- Timothy:
    - Updated the RawDataParser module to handle more possible raw data file setups. Worked on handling of single fields that are split up into multiple fileds within the raw data xml file. For example, if the TAX_ID of an incoming record is equal to 1234 than this may be stored in the xml as

```
<Identifier>

    <IdentifierType>SSN</IdentifierType>

    <IdentifierValue>1234<IdentifierValue>
```

</Identifier>

- Timothy set up our raw data parser to be able to handle these cases as well as more general cases in the XML.

Pending issues

- Logan:
  - Update the design diagram
  - Generate report for found inconsistencies
- Christopher:
  - Handling multiple reporters
  - Create more query builders for testing
- Camden:
  - Setup performance testing for CI
  - Refactor Parsing module
- Timothy:
  - Raw Data Parser Tests
  - Research deployment to AWS

Individual contributions

Team Member	Weekly Hours	Total Hours
Logan Heitz	18	62
Christopher Konopka	15	51.5
Camden Voigt	15	53
Timothy Rogers	16	58

Plan for coming week

- Logan:
    - Create updated block diagram for design
    - Generate report for found inconsistencies
- Christopher:
    - Handling multiple reporters
    - Create more query builders for testing
- Camden:
    - Setup performance testing for CI
    - Refactor Parsing module
- Timothy:
    - Raw Data Parser Tests
    - Research deployment to AWS

Summary of weekly advisor meeting

01/12/2018 Advisor Meeting

- Discussed the tasks that needed to be accomplished during this period including refactoring the exporter module into the storage module, updating the raw data parser to handle all incoming fields
- Talked about the possibility of having false negatives and the impact on the project. For now, we have suspended the use of hashing in order to eliminate this problem until we are able to get clarification from Kingland and test the performance of using just raw values.
- We discussed the potential need to have multiple central databases or tables to accommodate different kinds of reports and need to get clarification from Kingland on this. Currently our design is setup such that this would be easy to support.

01/19/2018 Advisor Meeting

- Discussed upcoming Kingland meeting that is setup for 01/31/2018. We will present what we have at this point and our progress from the last meeting and ask the questions that have come up since our last meeting. They will be providing us with several things, including inconsistency report example, that were discussed in our last meeting with them. We will need to discuss Amazon Web Service more with Kingland to see where

they are at in getting us a test instance and to figure out what features they want us to explore.

- Query Building needs to be moved out of the storage module into its own module to facilitate testing of different queries.
- The raw data parser is still a work in progress as the formatting of the file has made it difficult to parse through without having many hardcoded values in the module, which would not be extensible. For now, we will work to get the format we have working in the best way and discuss the incoming format with Kingland at the next meeting as they said it may be possible to send us a consistent format.

01/26/2018 Advisor Meeting

- Developed slides for presentation to Kingland next Wednesday.
- Discussed plans to investigate AWS and discuss our ideas with Kingland next Wednesday.
- Fixed some small issues in the codebase that affected the performance of the project.