sdmay18-27: Data Analytic Tools for Inconsistency Detection in Large Data Sets

Week 7 Report

October 21 - October 27

Team Members

Camden Voigt — *Technical Lead*
Christopher Konopka — *Communication Lead*
Logan Heitz — *Project Lead*
Timothy Rogers — *Quality Lead*

Summary of Progress this Report

During this week we continued with the implementation of the Data configuration parser as well as finished documentation for and began implementation of the Raw Data Parser. An early version of the matching configuration file was also created. Research into what XML parser we will use was also done.

Data configuration parser
For the data configuration parser this week we added in junit tests to verify the functionality of the parser. The tests will create a test configuration file and then parse it and verify that the correct values were recovered.

Raw Data Parser
Updated documentation by removing the database connection that is not needed and clarified how methods would work and what parameters they would use. Implementation was also started on this, the ParsedReport object has been created and the initial parser for XML was created using the SAX parser (more on this later).

Matching configuration file
An early version of this configuration file was created. The file will have an input path to the data as well as match-elements that specify which key-elements should be compared for inconsistency detection.

XML Parser
Since we will be parsing through rather large XML files (> 100 GB for daily reports) we knew we needed to use an efficient XML parser. Our original choice had been to use the Java DOM parser, however this parser loads in the entire DOM into memory at once. This can lead to memory requirements of roughly 3 times the size of the original XML file. This is far too much memory to use, instead we determined the SAX parser would be better for our requirements as it uses far less memory (does not load the DOM all at once) and is also faster than the DOM parser.

Pending Issues

Update Data configuration parser to use SAX parser (Logan Heitz)
Raw Data Parser Prototype (Christopher Konopka and Timothy Rogers)
Matching Configuration Parser Prototype (Logan Heitz)
Inconsistency Detector (Jackson Voigt)

Plans for Upcoming Reporting Period

Update Data configuration parser to use SAX parser (Logan Heitz)
Raw Data Parser Prototype (Christopher Konopka and Timothy Rogers)
- Store the information retrieved from the file (Christopher Konopka)
- Determine a way to figure out when a tuple ends while parsing (Timothy Rogers)

Matching Configuration Parser Prototype (Logan Heitz)
Inconsistency Detector (Jackson Voigt)

Individual Contributions

Team Member	Contribution	Weekly Hours	Total Hours
Camden Voigt	Updated documentation for Data configuration parser. Reviewed merge requests for Data configuration parser	2.5	18.5
Christopher Konopka	Updated documentation for Raw data parser.	2	14
Logan Heitz	Research into XML parsers and time/space savings. Implementation of tests for data configuration parser.	5	21.5
Timothy Rogers	Research into XML parsers and time/space savings. Updated documentation for Raw data parser. Prototype implementation of raw data parser	6	19