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

Week 6 Report
October 14 - October 20

Team Members

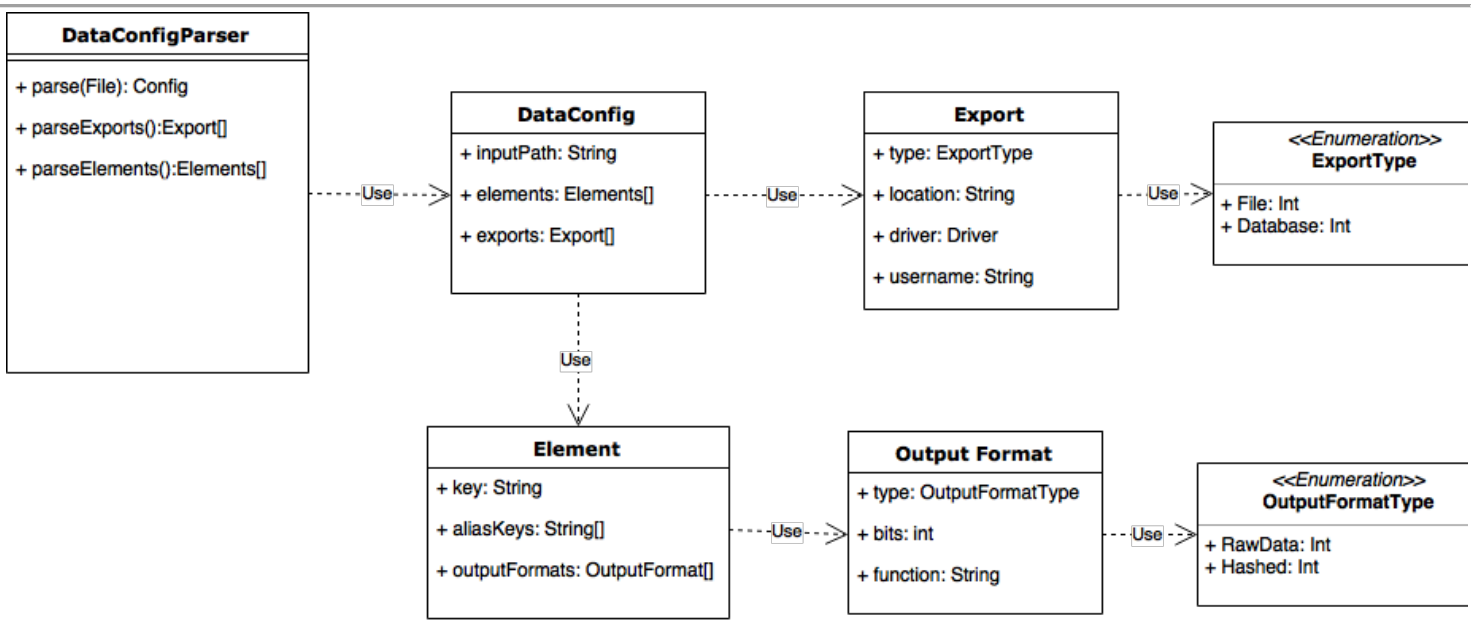
Camden Voigt — *Technical Lead*
Christopher Konopka — *Communication Lead*
Logan Heitz — *Project Lead*
Timothy Rogers — *Quality Lead*

Summary of Progress this Report

During this week we further developed the documentation for the raw data parser while we began coding of the data configuration parser as well as how matching will be done.

Raw Data Parser
For this file we discussed in a technical meeting with our advisor the initial documentation. It was determined that we do not need to store which specific report the tuples came from as knowing their id will be enough to reference back to the original data. Instead of inserting into the database at this stage we first want to pass the parsed data to the matcher that will check for inconsistencies before insertion. There is no need to keep raw data types, we can store things as strings. The storage key format for hashed values will be key_HashFunction. This will allow for us to test multiple different types of hashing at once without running into issues with having the same key. For now we will assume there is only one type of hash function and build our prototype accordingly.

Data Configuration Parser
While this used to be simply called the configuration parser we have determined that another configuration file will be needed to specify how matching should be down. So to differentiated the two we have renamed this the data configuration parser and the future one will be the matching configuration parser. We have developed a UML class diagram (included below) for this and begun the implementation of the different classes. So far everything has been coded and only the DataConfigParser is awaiting a merge request. The next step is to develop junit tests for this in order to verify that our code works properly.



Matching Configuration file
The matching configuration file will be similar to the data configuration file. This file will specify which columns needs to be matched to check for inconsistencies. A more formal outline of the configuration parser will be completed by the next meeting with our advisor by his graduate student working on this project.

Runtime
Our advisor would like us to get the runtime of all the functions we write so that we might find potential bottlenecks in the future if our performance needs improvement. This will be completed as code is finished before it is merged into the master branch of our project.

Pending Issues

Complete merge request for DataConfgParser (Camden Voigt)
Create JUNIT tests for DataConfigParser (Logan Heitz)
Update documentation of Raw Data Parser (Christopher Konopka and Timothy Rogers)
Update naming consistency between files and inform group and advisor (Logan Heitz)
Raw Data Parser Prototype (Christopher Konopka and Timothy Rogers)
Matching Configuration file Format
Matching Configuration Parser Prototype
Inconsistency Detector

Plans for Upcoming Reporting Period

Complete merge request for DataConfgParser (Camden Voigt)
Create JUNIT tests for DataConfigParser (Logan Heitz)
Update documentation of Raw Data Parser (Christopher Konopka and Timothy Rogers)
Update naming consistency between files and inform group and advisor (Logan Heitz)
Begin Raw Data Parser Prototype (Christopher Konopka and Timothy Rogers)

Individual Contributions

Team Member	Contribution	Weekly Hours	Total Hours
Camden Voigt	Implemented the DataConfig, Export, Element and OutputFormat objects.	3	16
Christopher Konopka	Presented documentation for the raw data parser to the group and advisor and asked clarifying questions for how the parser should be improved. Worked on updating documentation with findings from technical meeting and presentation.	2	12
Logan Heitz	Implemented the DataConfigParser that will use objects implemented by Camden Voigt to store data parsed from the data configuration file.	3	16.5
Timothy Rogers	Presented documentation for the raw data parser to the group and advisor and asked clarifying questions for how the parser should be improved. Worked on updating documentation with findings from technical meeting and presentation.	2	13