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

Week 4 Report
September 30 - October 6

Team Members

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

Summary of Progress this Report

Determined format for the configuration file that will be used by our parser to convert files from raw XML into various storage formats for testing. The configuration file will have the following outline

```
<?xml version="1.0" encoding="UTF-8"?>
<configuration>
    <!-- Input file location. If this is a folder, use all the files in the folder -->
    <inputfile>c://</inputfile>

    <!-- A list of output methods -->
    <outputs>
            <!-- Define different method base on formats -->
            <output type="database">
                    <dirver>mysql</dirver>
                    <location>localhost:3306</location>
                    <username>username</username>
                    <password>password</password>
            </output>
            <output type="file">
                    <location>c://</location>
            </output>
    </outputs>

    <!-- A list of key attributes we want to parse out -->
    <key-attributes>
            <!-- Define a attribute base on "key", "hash function" and "optional-names" -->
            <!-- If we don't want to hash this attribute, we can leave hash-function to empty -->
            <attribute key="ssn">
                    <!-- In the format we define the output format to store -->
                    <!-- We have raw, binary, and we can use name of hash function to specify hash value -->
                    <output-format>binary</output-format>
                    <output-format>md5</output-format>
            </attribute>
            <!-- Optional name gives a list of other key for same value -->
            <attribute key="name" optional-names="firstname,lastname">
                    <output-format>raw</output-format>
                    <output-format>binary</output-format>
```

```
                <output-format>sha-1</output-format>
            </attribute>
    </key-attributes>
</configuration>
```

This format has three major sections. The first is the input location, this can be either a file or a folder. If a folder is specified we will get all files in the folder. This will allow us to get multiple reports at the same time. The second section is the output. This will specify the different outputs that we will generate and other important information such as usernames and passwords for a database. Finally there will be a key attribute section that will specify which attributes we want to get from the raw data and how they should be stored in the database i.e. raw data, hashing, etc.

Pending Issues

Outline for a configuration parser that will take the configuration file and build a configuration object in C#.
Outline for a raw data parser that will use a configuration file to parse the reports and convert them based on the output format.
Creation of design documentation V1 for the project
Prototype of configuration parser
Prototype of raw data parser

Plans for Upcoming Reporting Period

Create outline of configuration parser and creation of design document V1

Individual Contributions

Team Member	Contribution	Weekly Hours	Total Hours
Camden Voigt	Updated proposed solution and block diagram with new information. Worked on the formatting of the configuration file.	2	9
Christopher Konopka	Updated goals and purpose on the project plan. Worked on the formatting of the configuration file.	1.5	7
Logan Heitz	Updated project schedule with new information. Worked on the formatting of the configuration file.	2	9.5
Timothy Rogers	Updated team website with new technical details on the project. Worked on the formatting of the configuration file.	1	8