EE/CprE/SE 492 – sdmay18-27

Data Analytic Tools for Inconsistency Detection in Large Data Sets

Week 2 Report

01/27/2018 – 02/09/2018

Client/Advisor: Kingland Systems, Ying Cai

Team Members/Role:

Logan Heitz – Project Lead

Christopher Konopka – Communication Lead

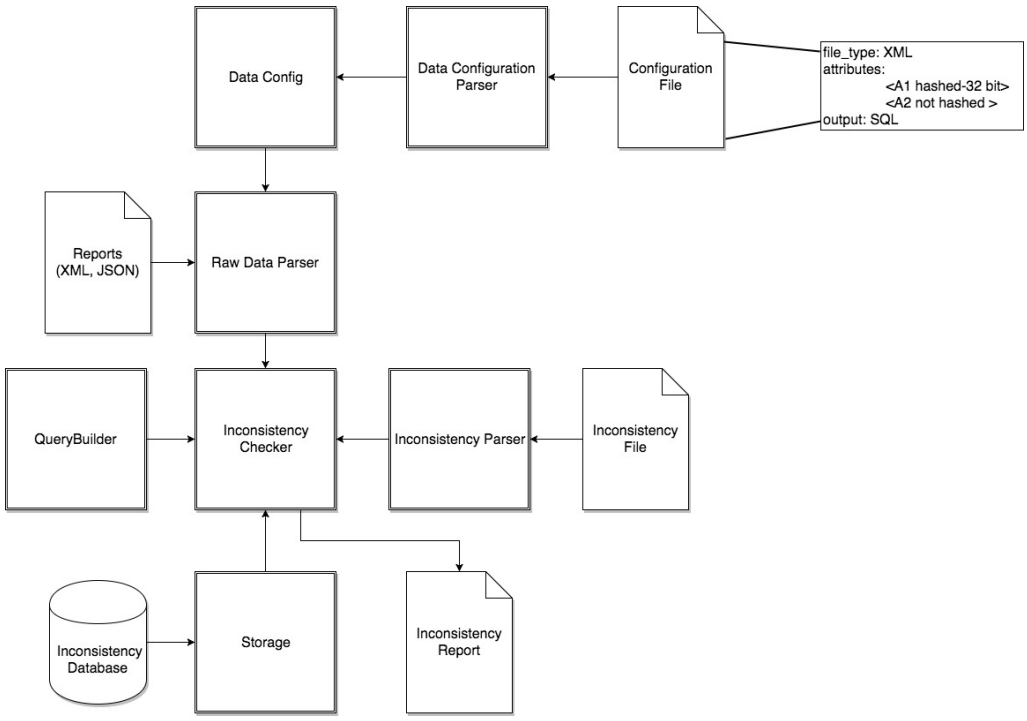
Camden Voigt – Technical Lead

Timothy Rogers – Quality Lead

Weekly Summary

Past week accomplishments

- Logan:
    - Refactored the configuration file and the configuration parser to have a single storage type specified as opposed to a list of storage types. While the original intent was to have multiple storage types to facilitate testing of different types, we decided that only having 1 specified will still allow for easy testing and will be what is desired in the final product.
    - Updated the block diagram for the project to illustrate new changes to the project. This block diagram was used in a presentation to Kingland to give them an update on our progress. Below is the updated block diagram.



  - Removed the hashing module from the project after discussing the potential issues with Kingland.
  - Removed the output type from the key element list as all data will be stored as raw data for inconsistency checking.
  - Developed documentation for REST API that can be used in the UI mockup that we will create for Kingland.
  - Developed new database documentation with proposal to separate out tables by reporters to shrink table size and provide better ability for parallelization.
- Christopher:
  - Fixed critical with inconsistencies fields having incorrect names in reporting
  - Added in the ability to check for reporter Id when checking inconsistencies and to classify the inconsistency as appropriate based on the reporter Id.
- Camden:
  - Setup performance testing for our project when merging into master. These tests will launch a full run of the project when the develop branch is merged into master so that we can collected benchmarks for the project as development is in progress.

    - Prototyped multithreading for the project and collected performance metrics.
    - Researched deployment on AWS
- Timothy:
    - Created updated tests for the raw data parser using mockito.
    - Added reporter Id as a parsable filed for the raw data parser.
    - Research into using HighCharts to display inconsistency results as discussed during the meeting with Kingland on 01/31/2018.

Pending issues

- Logan:
    - Implement new database setup
- Christopher:
    - Integrate new database setup with the inconsistency checker
- Camden:
    - Begin setup of UI using HighCharts
    - Setup REST API
- Timothy:
    - Continue setup of multithreading using new design

Individual contributions

Team Member	Weekly Hours	Total Hours
Logan Heitz	17	78
Christopher Konopka	15	66.5
Camden Voigt	17	70
Timothy Rogers	15	73

Plan for coming week

- Logan:
    - Implement new database setup
- Christopher:

    - Integrate new database setup with the inconsistency checker
- Camden:
    - Begin setup of UI using HighCharts
    - Setup REST API
- Timothy:
    - Continue setup of multithreading using new design

Summary of weekly advisor meeting

01/31/2018 Kingland Meeting

- AWS
    - We will probably start with a single instance to test the performance benefits of it over our current machine.
    - M4.2xlarge instance is what Kingland has typically used in the past.
    - Could look at doing a memory optimized instance.
    - Amazon Aurora is a possibility to help relational database performance.
    - Kingland is still working on getting the instance setup so we have the proper access.
- Hashing
    - Won't improve our space saving by very much and Kingland is fine with us removing the hashing.
    - They prefer that we avoid the false negatives as well.
- Inconsistency list
    - The sample we have is representative of the whole list, so we shouldn't need any more for tests.
- HighCharts
    - This is what Kingland uses for displaying information.
    - We can start to play around with this to provide a proof of concept for how our information can be displayed.
- Raw Data
    - Currently the data format is not set in stone, but if nothing changes we can treat the XML format we have currently as the only valid format.
    - Kingland will try to provide us with an XML schema for the raw data.

02/02/2018 Advisor Meeting

- Discussed direction going forward following the information we learned at the Kingland meeting this week
- Decided to have Camden look into doing multi-threading

- Decided to have Logan look into storing the inconsistencies in a database table and getting documentation for a REST API started.
- Timothy will look into using HighCharts so that we can mock up a UI for Kingland.

02/09/2018 Advisor Meeting

- Decided to move forward with the proposed database design revision and REST API design
- Reassigned Camden to working on the UI and REST API due to previous experience
- We will try to do some more tests for the multithreading implementation to get better time measures.