СИСТЕМА ПОИСКА И АГРЕГАЦИИ ДАННЫХ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Цубер Е. Г.

Осипович В. С. – к-т техн. наук, доцент

Целью работы является разработка системы поиска и агрегации данных, предназначенной для импорта данных из различных источников, их обработка и миграции в другие системы. Данная система позволяет интегрироваться практически с любыми источниками получения данных и точками назначения. В данной системе предоставлен гибкий механизм для обработки данных (обработка различных форматов и др), а также возможность реагировать на исключительные ситуации, например, отправка уведомлений по почте при обнаружении важных данных при обработке и др. В данной системе реализован механизм для аналитики состояния системы, анализа загруженности, статистики по числу обработанных документов и др.

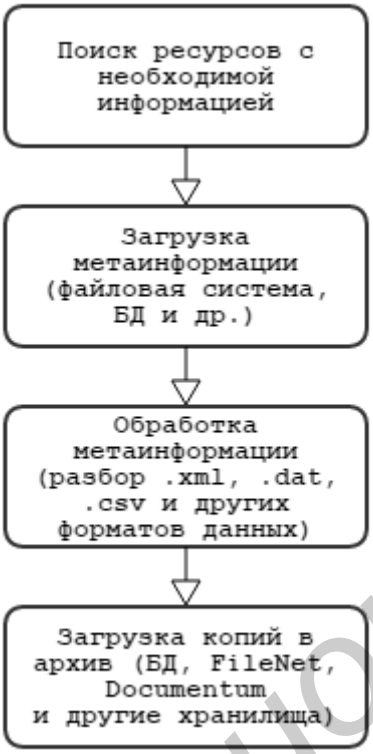

Рис. 1 – Общая архитектура системы

Основные функции системы – это загрузка и обработка данных. Загрузчики срабатывают в определённые интервалы времени и производят поиск новых данных для обработки. При обнаружении, каждый найденный блок данных ставится в очередь на обработку для последующей отладки обработчиком. Также для блока данных можно указать тип обработчика, которым он будет обработан. Из очереди блоки данных начинают поступать на обработку порциями через определённые интервалы. Таким образом, достигается равномерное и последовательное поступление данных на обработку. В обработчиках инкапсулируется логика по обработке блоков данных.

Для описания логики работы загрузчика и обработчика предоставляется возможность использовать скриптовый язык, за счёт этого, можно реализовать абсолютно любую логику по загрузке и обработке данных. Также реализована возможность интегрировать загрузчик и обработчик с различными системами, например, для получения справочной информации или отправки уведомлений. При написании скрипта по обработке данных, предоставляется гибкое API для разбора различных форматов (json, xml и др). Ключевой особенностью работы обработчиков является возможность из запуска в несколько потоков. Таким образом, за счёт одновременной работы нескольких обработчиков, можно увеличить пропускную способность системы во много раз (вертикальное масштабирование).

В системе присутствует возможность в реальном времени следить за процессом обнаружения и обработки новых данных. Также предоставляется удобный механизм для поиска блоков данных. Можно найти блоки, которые были обнаружены в определённый интервал времени, либо с определённым статусом и просмотреть результаты их обработки. В случае возникновения ошибок при обработке, можно доработать обработчик, и из данного интерфейса отправить блоки данных на повторную обработку, таким образом, данные не теряются, если они не были обработаны с первого раза.

В разработанной системе имеется страница со статистикой. Она предназначена для просмотра состояния системы в реальном времени. Отображается информация о текущей загруженности обработчиков, числе найденных и обработанных блоков данных.

В ходе работы создана система поиска и агрегации данных на Java, которая обеспечивает сбор информации из различных источников, их обработку и сохранение в определённом формате.

Список использованных источников:

1. Эккель Б. Философия Java. Библиотека программиста/Б. Эккель. — СПб Питер, 2012. – 640 с.:ил.
2. Паттерны проектирования / Б. Бейтс, К. Сьерра, Э. Фримен. – СПб.: Питер, 2011. – 656 с.
3. Программирование на Java для детей, родителей, бабушек и дедушек – Электронные данные. – Режим доступа: 0-9718439-5-3.pdf
4. The XML C parser and toolkit of Gnome [Электронный ресурс http://www.xmlsoft.org/]