**САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ**

**УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ,   
МЕХАНИКИ И ОПТИКИ**

Факультет информационных технологий и программирования

Кафедра информационных систем

Направление подготовки (специальность) 09.04.02

**О Т Ч Е Т**

**о производственной практике**

(наименование практики)

**Тема задания: Улучшение и внесение изменений в высоконагруженную платформу по хранению, обработке и предоставлению информации с возможностью детальной настройки правил обработки входных и выходных данных**

**Студент \_Трофимов В.А.\_ \_\_\_\_\_\_\_М4106\_\_\_\_\_\_\_**

(Фамилия И.О.) номер группы

**Руководитель практики от организации:** Гедзберг М.Ю., ООО "Люксофт Профешнл", старший

руководитель программы

(Фамилия И.О., должность и место работы)

**Ответственный за практику от университета:** Зубок Д.А., доц. каф. ИС

(Фамилия И.О., должность)

**Практика пройдена с оценкой \_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Подписи членов комиссии**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_( )**

(подпись)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_( )**

(подпись)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_( )**

(подпись)

**Дата \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Санкт-Петербург**

**2017**

Задачей практики является улучшение и внесение изменений в высоконагруженную платформу по хранению, обработке и предоставлению информации с возможностью детальной настройки правил обработки входных и выходных данных.

Задачи на период практики:

* Создание модуля по автоматизации создания ежедневных бэкапов данных, хранящихся в системе, с удалением из них конфиденциальных данных для последующего анализа Модификация и написание утилит для автоматизации развертывания платформы на тестовые окружения
* Создание набора скриптов для работы со структурированными CSV файлами большого объема, обработка которых средствами Excel или другими процессорами электронных таблиц невозможна
* Разработка платформы для нагрузочного и функционального тестирования платформы
* Разработка платформы по автоматизированному перемещению обработанных бэкапов с данными на тестовые окружения с учетом ограничений безопасности и объема дискового пространства узлов-посредников передачи данных

## Описание рассматриваемой платформы

Платформа представляет из себя высоконагруженное распределенное приложение, организующее автоматический забор данных из заранее сконфигурированных источников, проверку соответствию формата данных, трансформацию и их сохранение во внутреннюю базу данных и высокопроизводительный распределенный кэш в структурированном формате, позволяющем эффективно осуществлять поиск по хранимой информации. Впоследствии данные отдаются конечным пользователям через REST-интерфейс.

На данный момент поддерживается забор данных из CSV файлов и из реляционных баз данных через JDBC соединение.

Приложение является полностью распределенным — для изменения размера кластера необходимо указать в файлах настроек соответствующего узла информацию о всех узлах кластера.

Основные слои и компоненты платформы изображены на рисунке.

## 

Основные слои и компоненты:

* External Feeds – внешние источники данных, из которых платформа выкачивает данные по расписанию
* External Services – внешние системы для обеспечения идентификации, аутентификации и авторизации пользователей в системе
* Data Processing Service – управление забором данных из внешних источников и ее обработка
* Configuration Management Services – хранит информацию о забираемой и отдаваемой информации, настройках соединений с источниками, расписаниях и правилах обработки данных
* Rule Engine – распределенный сервис по выполнению правил над поступающей информацией
* REST Services – предоставляет внешним пользователям доступ к хранимым данным
* Data Layer – слой хранения данных и настроек системы

Основные технологии и продукты, используемые в платформе:

* Hazelcast: распределенный кэш; предоставляет масштабируемый, высоконагруженный и отказоустойчивый доступ к данным
* Drools: ядро выполнения правил по обработке данных, позволяет с помощью высокоуровневых операций описывать принципы, по которым будут обработаны входные и выходные данные
* Spring: IoC контейнер
* Camel: координатор; позволяет управлять потоком данных и управления декларативно, используя язык разметки xml
* Tomcat: контейнер для веб-приложений
* JMX: Протокол управления и мониторинга для Java
* JDBC: протокол взаимодействия с БД
* Oracle DB: реляционная СУБД

## Работы, выполненные в рамках практики

В рамках практики были выполнены следующие работы:

* Создание модуля по автоматизации создания ежедневных бэкапов данных, хранящихся в системе, с удалением из них конфиденциальных данных для последующего анализа Модификация и написание утилит для автоматизации развертывания платформы на тестовые окружения
* Создание набора скриптов для работы со структурированными CSV файлами большого объема, обработка которых средствами Excel или другими процессорами электронных таблиц невозможна
* Разработка платформы для нагрузочного и функционального тестирования платформы
* Разработка платформы по автоматизированному перемещению обработанных бэкапов с данными на тестовые окружения с учетом ограничений безопасности и объема дискового пространства узлов-посредников передачи данных

## Создание модуля по автоматизации создания ежедневных бэкапов данных, хранящихся в системе, с удалением из них конфиденциальных данных для последующего анализа Модификация и написание утилит для автоматизации развертывания платформы на тестовые окружения

Для более эффективного анализа и внесения изменений в продукт требуется анализ реальных данных, загружаемых в систему. Для этого необходим анализ данных, которые хранятся на окружении, используемым заказчиком. Однако, разглашение или получение конфиденциальных данных в окружение разработчиков недопустимо, вследствие чего был разработан модуль, который в автоматизированном режиме после каждой новой загрузки данных в систему создает архив со всеми не конфиденциальными данными в директории, доступной для доступа разработчикам.

## Создание набора скриптов для работы со структурированными CSV файлами большого объема, обработка которых средствами Excel или другими процессорами электронных таблиц невозможна

Частью процесса анализа является анализ тестовых и реальных данных, которые сохраняются в платформу, а также входные выходные данные, полученные в результате различных запросов к платформе. Зачастую объем CSV и JSON файлов, подлежащих обработке, превышает 2ГБ, что является предельно допустимым объемом файла для обработки его с помощью 32-разрядного Excel. В дополнение, обработка данных большего объема даже с использованием 64-разрядного Excel является крайне затруднительной задачей ввиду крайне большого времени отклика интерфейса.

Для упрощения работы с такими данными был разработан ряд утилит для решения различных задач, в том числе:

* Преобразование данных из формата CSV в набор запросов с указанными параметрами
* Преобразование ответов от платформы в один файл формата CSV
* Фильтрация, поиск и преобразование CSV-файлов
* Пересечение CSV файлов для выполнения операций, схожих по принципу с SQL JOIN
* Вычисление различий в нескольких CSV файлах, имеющих одинаковую структуру

## Разработка платформы для нагрузочного и функционального тестирования платформы

Используемые ранее платформы и утилиты для функционального и нагрузочного тестирования платформы потребляли недопустимо большой объем оперативной памяти. Аналогичные утилиты, позволяющие осуществлять те же операции с использованием допустимого объема оперативной памяти, не являются допустимыми к использованию в рамках проекта. В связи с этим было принято решение о разработке собственной высокоэффективной платформы на языке Python для выполнения требуемых задач.

## Разработка платформы по автоматизированному перемещению обработанных бэкапов с данными на тестовые окружения с учетом ограничений безопасности и объема дискового пространства узлов-посредников передачи данных

Архивы с не конфиденциальными данными, хранящиеся на окружении, используемом заказчиком, требуют переноса на окружения разработчиков для доступа к ним. Однако, такая передача возможна только через специализированный узел-посредник, аутентификация на котором осуществляется исключительно с использованием ввода пароля, и на котором доступно всего около 5ГБ дискового пространства. Процесс по переносу таких архивов выполнялся ежедневно вручную самими разработчиками. В рамках практики была выполнена автоматизация этого процесса, поиск подходящих библиотек и инструментов, инкапсулирующих работу с таким узлом-посредником без компрометации персональных логинов и паролей разработчиков, разработка набора скриптов для отслеживания изменений и появления архивов, готовых к передаче, с учетом того, что один или несколько переданных архивов могут быть больше, чем доступное на данный момент свободное дисковое пространство.

## Вывод

По результатам практики были выполнены все поставленные задачи в полном объеме.