Министерство образования Республики Беларусь
Учреждение образования
«Белорусский государственный университет
информатики и радиоэлектроники»

УДК 004.624

ЗАНЬКО
Роман Владимирович

ETL-СИСТЕМА ДЛЯ ПРЕДОСТАВЛЕНИЯ СТАТИСТИЧЕСКИХ ДАННЫХ ИЗ АИС «ЭЛЕКТРОННЫЙ РЕЦЕПТ» В РАЗРЕЗЕ НОМЕНКЛАТУРЫ ЛЕКАРСТВЕННЫХ СРЕДСТВ

АВТОРЕФЕРАТ
диссертации на соискание степени
магистра информатики и вычислительной техники

по специальности 1-40 81 04 – Обработка больших объемов информации

Научный руководитель
Красько О.В.
к.т.н., доцент

Минск 2020

Работа выполнена на кафедре информатики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Научный руководитель: **КРАСЬКО Ольга Владимировна,**
кандидат технических наук, доцент, ведущий научный сотрудник лаборатории №212 государственного научного учреждения «Объединенный институт проблем информатики Национальной академии наук Беларуси»

Рецензент: **СТАРОВОЙТОВА Татьяна Феликсовна,**
кандидат экономических наук, доцент кафедры управления информационными ресурсами учреждения образования «Академия управления при Президенте Республики Беларусь»

Защита диссертации состоится «23» июня 2020 г. года в $10^{\underline{00}}$ часов на заседании Государственной экзаменационной комиссии по защите магистерских диссертаций в учреждении образования «Белорусский государственный университет информатики и радиоэлектроники» по адресу: 220013, Минск, ул. Гикало, 9, копр. 4, ауд. 112, тел. 293-85-91, e-mail: inform@bsuir.by

С диссертацией можно ознакомиться в библиотеке учреждения образования «Белорусский государственный университет информатики и радиоэлектроники».

ВВЕДЕНИЕ

На сегодняшний момент аналитические системы, системы искусственного интеллекта для прогноза и принятия решений на основе Big Data становятся неотъемлемой частью во многих сферах, таких как экономика, здравоохранение, образование и производство.

Сейчас, многие компании используют Data Driven подход для принятия решений. Главный постулат данной методологии - решения нужно принимать, опираясь на анализ цифр, а не интуицию и личный опыт. Подход подразумевает, что нужно понимать данные и уметь строить прогнозы на их основе. То есть на этапе принятия решения должно быть понимание, на что оно повлияет, что нужно изменить, какого результата можно добиться.

Ценность, надежность и достоверность знаний, полученных в результате интеллектуального анализа данных, зависит не только от эффективности используемых аналитических методов и алгоритмов, но и от того, насколько правильно подобраны и подготовлены исходные данные для анализа.

Поэтому, прежде чем приступать к анализу данных, необходимо выполнить ряд манипуляций с данными, цель которых — доведение данных до определенного уровня качества и информативности, а также организовать их интегрированное хранение в структурах, обеспечивающих их целостность, непротиворечивость, высокую скорость и гибкость выполнения аналитических запросов.

Одним из инструментов решения данных задач является процесс ETL. Данный процесс представляет собой комплекс операций, реализующих процесс переноса первичных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных. Является составной частью этапа консолидации данных в анализе данных.

В рамках мероприятия 21 «Создание полномасштабной системы обращения электронных рецептов в Республике Беларусь с использованием электронной цифровой подписи» Государственной программы развития цифровой экономики и информационного общества на 2016 – 2020 годы создан опытный образец АИС "Электронный рецепт". За весь период работы АИС ЭР выписано почти 12 млн рецептов. В настоящий момент стоит задача получить статистическую и аналитическую информацию из АИС ЭР для определения затрат на медикаменты по определенным нозологиям и планирования закупок лекарственных средств. Поэтому для выполнения поставленной задачи нужно построить информационную систему, которая предоставляет качественные медицинские данные для построения необходимой аналитической отчетности и прогноза.

Настоящая работа посвящена ETL-системе подготовки больших объемов данных медицинской направленности в интересах здравоохранения Республики Беларусь.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы является разработка процессов выгрузки информации из источников, ее последующей трансформации и загрузки в хранилище данных, а также автоматизация данных процессов, для повышения качества медицинских данных для построения аналитических и статистических отчетов, и визуальной аналитики.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Провести обзор и анализ накопленного опыта в области построения ETL-систем, а также рассмотреть инструменты построения таких систем.
2. Провести анализ данных и метаинформации в источниках и хранилище данных и выявить основные проблемы интеграции.
3. Разработать процессы для интеграции информации, и произвести их автоматизацию с помощью разработки ETL-системы.
4. Проанализировать результат работы системы и дать оценку качеству данных предоставляемых ETL-системой.

Объектом исследования являются разнородные медицинские данные и методы их извлечения и трансформации для различных целей.

Предметом исследования является система интеграции данных между источниками и хранилищем.

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность использования накопленной информации в АИС ЭР для предоставления аналитических и статистических данных, а также визуальной аналитики в разрезе номенклатуры лекарственных средств. Это позволит строить прогнозы и определять затраты на медикаменты по определенным нозологиям и планировать закупки лекарственных средств Оперативный и многофункциональный анализ больших объемов данных расширит функциональные возможности для поддержки принятия решений.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

Работа выполнялась в соответствии с мероприятием 21 «Создание полномасштабной системы обращения электронных рецептов в Республике Беларусь с использованием электронной цифровой подписи» подпрограммы 3 «Цифровая трансформация» Государственной программы развития цифровой экономики и информационного общества на 2016 – 2020 годы, утвержденной постановлением Совета Министров Республики Беларусь от 23 марта 2016 г. № 235. А также договор между Министерством здравоохранения Республики Беларусь и государственным учреждением «Республиканский научно-практический центр медицинских технологий, информатизации, управления и экономики здравоохранения» от 14 декабря 2016 г. № ЦТ21-ЭР.

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя О. В. Красько, заключается в формулировке целей и задач исследования.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, трёх глав, заключения, списка использованных источников и приложений. В первой главе представлен анализ ETL-процессов, обзор критериев качества данных и существующих инструментов для построения ETL-системы. Вторая глава посвящена анализу данных и метаинформации в ХД, АИС ЭР и других источниках, а также в ней рассматриваются проблемы интеграции и способы их решения. В третьей главе представлены задачи и требования перед системой предложен инструмент построения ETL-системы, описаны процессы и их разработка, а также проведен анализ результат работы системы и дана оценка качеству данных.

Общий объем работы составляет 63 страницы, из которых основного текста – 42 страниц, 11 рисунков на 7 страницах, 5 таблиц на 10 страницах, список использованных источников из 25 наименований на 2 страницах и 3 приложения на 21 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** рассмотрено современное состояние проблемы повышения качества данных для прогнозирования и принятия решений.

В **общей характеристике работы** показана актуальность проводимых исследований, степень разработанности проблемы, сформулированы цель и задачи диссертации, обозначена область исследований, научная (теоретическая и практическая) значимость исследований.

В первой главе проведен анализ ETL и его основных функций, и задач. Приведен обзор критериев оценки качества данных, а также рассмотрены существующие инструменты построения ETL-систем.

Из анализа ETL следует, что основными его функциями являются извлечение данных из источников, их преобразование и дальнейшая загрузка в хранилище данных. Перемещение данных в процессе ETL можно разбить на последовательность процедур, представленных следующей функциональной схемой (рисунок 1).

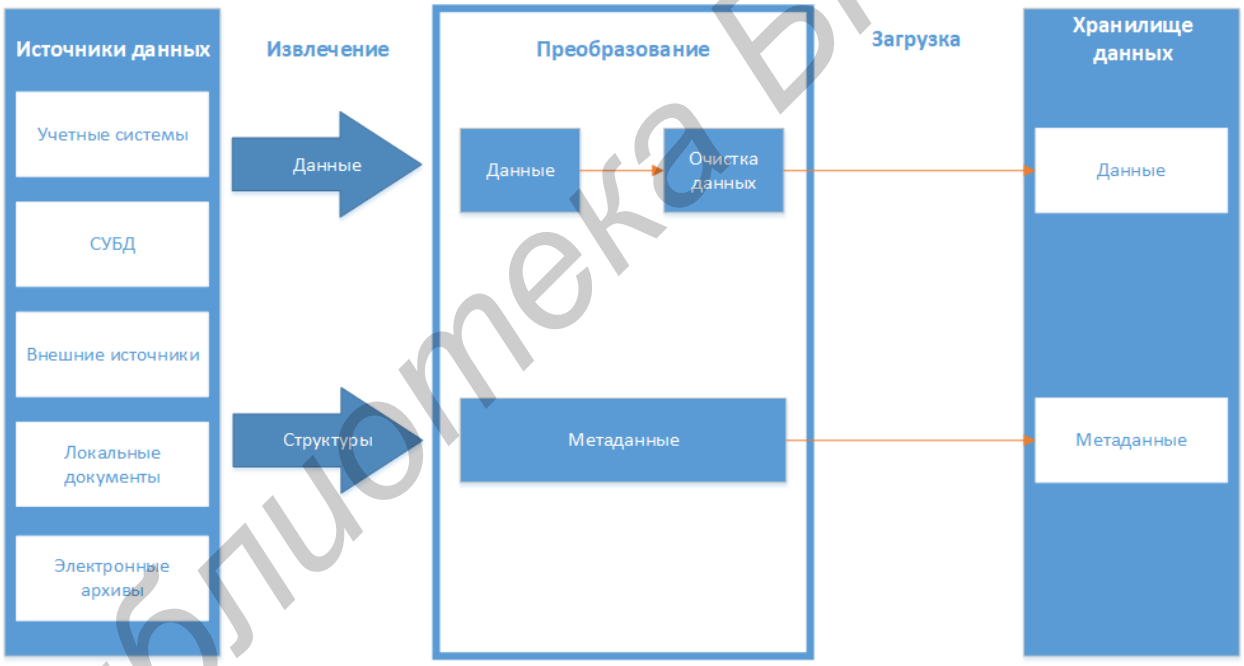

Рисунок 1 – Функциональная схема ETL-процесса

Проанализированы критерии оценки качества данных. Качество данных является обобщённым понятием, отражающее степень их пригодности к решению определённой задачи. В соответствии со стандартом ISO 9000:2015 основными критериями качества являются полнота, достоверность, точность, согласованность, доступность и своевременность.

Рассмотрены существующие инструменты построения ETL-систем. Инструменты ETL экономят время и деньги при разработке хранилища данных, устраняя необходимость в «ручном кодировании». «Ручное кодирование» по-

прежнему является наиболее распространенным способом интеграции данных сегодня.

Во второй главе проанализирована метаинформация хранилища данных для построения отчётов, АИС «Электронный рецепт» и справочников РУП «Белфармация». Также был произведен анализ проблем интеграции данных между источниками и хранилищем.

АИС «Электронный рецепт» предназначена для реализации технологии обращения электронных рецептов в здравоохранении Республики Беларусь и представляет собой централизованную систему электронной выписки и отпуска лекарственных средств при лечении в амбулаторных и стационарных условиях, включая льготное лекарственное обеспечение.

Хранилище данных для предоставления статистических данных из АИС «Электронный рецепт» в разрезе номенклатуры лекарственных средств использует схему «снежинка» с одной таблицей фактов, которая находится в схеме «Fact» и шестью таблицами измерения, которые находятся в схеме «Dim» (рисунок 2)

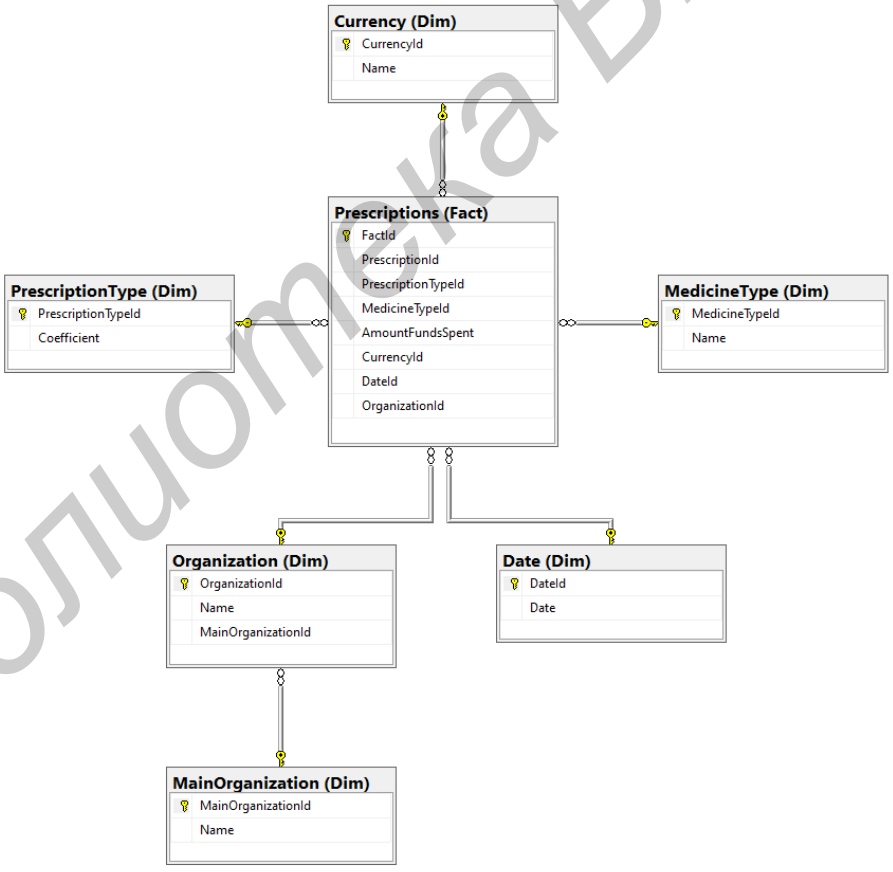

Рисунок 2 – Архитектура ХД для предоставления статистических данных из АИС “Электронный рецепт” в разрезе номенклатуры лекарственных средств

АИС ЭР использует стандарт Fast Healthcare Interoperability Resources для хранения и обмена данными. FHIR — стандарт обмена медицинской информацией. Стандарт описывает форматы медицинских данных и обмен

этими данными через REST API. FHIR является торговой маркой некоммерческой организации HL7 (Health Level Seven International).

Для работы АИС ЭР используются следующие ресурсы FHIR: *Patient*, *MedicationPrescription*, *MedicationDispense*, *Claim Practitioner*, *Organization*.

Для предоставления статистических данных из АИС «Электронный рецепт» в разрезе номенклатуры лекарственных средств используются справочники РУП «Белфармация». Данная организация предоставляет следующие справочники в формате XML: производители лекарственных средств, лекарственная форма ЛС, дозировка ЛС, единицы измерения ЛС, международное наименование ЛС и номенклатуры ЛС.

При анализе метаданных источников информации и ХД для предоставления статистических данных в разрезе номенклатуры лекарственных средств выявлены следующие проблемы:

1. АИС ЭР в системе интеграции данных является OLTP-системой. Это означает, что система интеграции данных должна создавать минимальную нагрузку на базу данных АИС ЭР, так как это может повлиять на скорость обработки транзакций и вследствие ухудшить качество предоставляемых услуг. На рисунке 3 и 4 представлены графики нагрузки на систему.

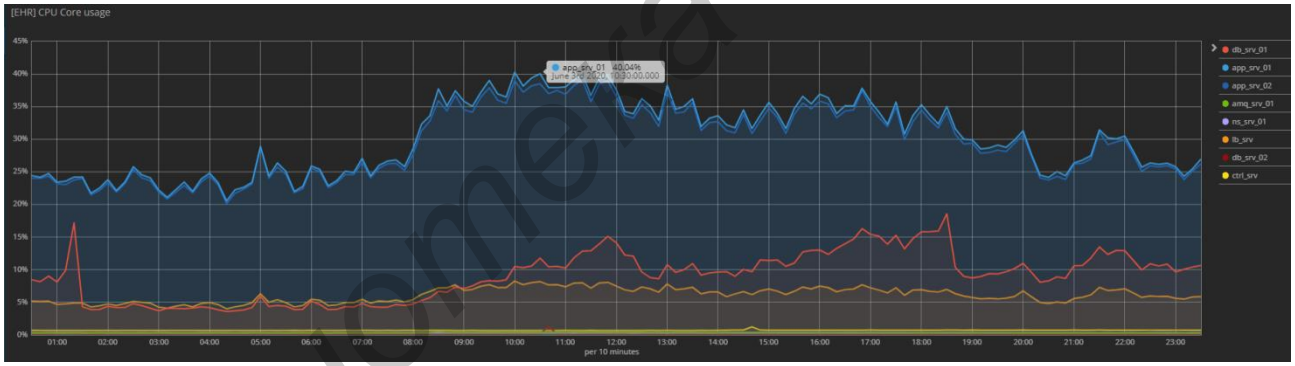

Рисунок 3– График нагрузки ЦП

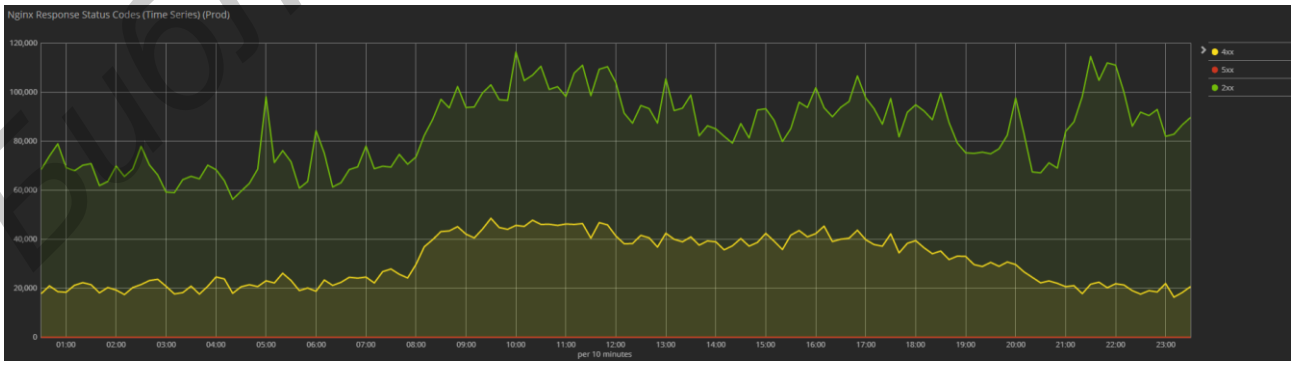

Рисунок 4 – График количества ответов АИС ЭР

2. АИС ЭР использует ресурсы FHIR для взаимодействия. Все ресурсы хранятся в БД в формате JSON. В результате это усложняет процедуру

извлечения данных, так как необходимо использовать дополнительные инструменты для поиска необходимых полей и их значения в JSON файле.

3. Для заполнения информации в ХД для предоставления статистических данных в разрезе номенклатуры лекарственных средств, нужно извлекать данные из нескольких FHIR-ресурсов. Это будет влиять на скорость процедуры извлечения, так как нужно проводить объединение этих данных. Операция объединения является трудоёмкой, особенно если объём данных большой.

4. Поскольку ресурсы FHIR хранятся в JSON, справочники РУП «Белфармация» в XML, а ХД для предоставления статистических данных в разрезе номенклатуры лекарственных средств хранит данные в реляционном виде с разными типами данных для каждого поля, необходимо будет проводить операции конвертации данных.

5. РУП «Белфармация» предоставляет справочники в формате XML. Как результат необходимо использовать инструменты для поиска необходимых полей и их значения в XML файле.

В третьей главе представлены задачи и требования к ETL-системе и описана разработка ETL-процессов. Также представлено обоснование выбора инструмента разработки ETL-системы и дана оценка качеству данных предоставляемых системой.

После анализа проблем интеграции и текущих возможностей АИС были выставлены такие основные требования: система должна быть кроссплатформенная, должна быть предусмотрена возможность выгрузки данных из АИС ЭР за указанный пользователем период, справочники для таблиц фактов должны поддерживаться в актуальном состоянии, нагрузка на АИС ЭР должна быть минимальной и все операции в данной системе должны проводится в выделенное окно времени, когда количество операций низкое.

Для разработки ETL-системы был выбран инструмент Microsoft Integration Services. Данный инструмент обладает необходимыми функциями для выполнения поставленных задач. Также есть необходимая документация по установке, конфигурации и разработке в Microsoft Integration Services, что ускоряет и упрощаете процесс разработки системы. Основным недостатком данной платформы является то, что она платная.

Для проведений операций над данными перед их непосредственные загрузки в ХД, была создана схема *Integration*, которая содержит таблицы промежуточной области.

Для выполнения задач системы были созданы 5 пакетов, каждый из которых выполняет конкретную задачу:

1. *Extract*. Задачей данного пакета извлечь все необходимые данные из источников и загрузить в промежуточную область.

2. *Transformation*. Пакет выполняет все необходимые трансформации над информацией перед загрузкой в ХД.

3. *LoadDim*. Главная цель пакета – загрузить данные в таблицы измерений.

4. *LoadFact*. Задачей пакета является загрузка информации в таблицу фактов.

5. *Main*. Является агрегирующим пакетом, который запускает на выполнение указанные выше пакеты, а также процедуру создания резервной копии и отправки сообщения на электронную почту.

На основе результатов работы системы была дана оценка качеству данных, которые предоставляет данная система.

Индекс стандартизированности данных предоставляемых системой является 0,875. Чем выше индекс стандартизированности, тем больше ХД соответствует нормам и стандартам, а данные из него могут использоваться для сравнения с аналогичными данными других регионов (стран) и быть пригодными для межцентровых исследований

Индекс своевременности информации составляет 30, что говорит об избыточности. Однако это обусловлено тем, чтобы уменьшить нагрузку на систему и увеличить качество предоставляемых услуг.

Индекс полноты данных в ХД составляет 1, т.к. данные с отсутствующими атрибутами не заносятся в ХД и не могут использоваться как достоверными. Данные которые являются не полными, должны рассматриваться отдельно в анализе и дальнейшем принятии решении на основе их.

Индекс востребованности для всех атрибутов составляет 1, поскольку все атрибуты учувствуют в аналитической отчётности.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Проведен обзор и анализ накопленного опыта в области построения ETL-систем. Также рассмотрены инструменты построения интеграциных систем. В качестве инструмента автоматизации ETL-процессов был выбран Microsoft Integration Services.

2. Проведен анализ данных и метаинформации в источниках и хранилище данных и выявлены основные проблем интеграции между АИС ЭР, справочниками РУП «Белфармация» и ХД для предоставления статистических данных в разрезе номенклатуры лекарственных средств. Для всех представленных проблем были предложены способы их решения на основе накопленного опыта в области построения ETL-систем

3. Разработаны и автоматизированы необходимые процессы для интеграции данных и предоставления их на высоком уровне качества. Основными процессами являются: выгрузка данных из всех источников, их очистка и трансформация, загрузка данных в таблицы измерений и фактов.

4. Проанализированы результат работы системы и дана оценка качеству данных предоставляемых ETL-системой. Индекс стандартизированности составляет 0,875. Индекс своевременности – 30. Индекс полноты данных в ХД составляет 1. Индекс востребованности для всех атрибутов – 1. На основе этого можно утверждать, что система предоставляет данные на достаточно высоком уровне качества, что бы на основе их можно было проводить аналитическую и статистическую работу.

.