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

УДК 004.428

Сафронова
Евгения Геннадьевна

Программный модуль хранения и обработки финансовой информации

АВТОРЕФЕРАТ
диссертации на соискание степени магистра

по специальности 1-40 80 04 – Информатика и технологии программирования

Научный руководитель
Пилецкий И. И.
к. ф-м. н., доцент

Минск 2022

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

Научный руководитель: **Пилецкий Иван Иванович,**
кандидат физико-математических наук, доцент кафедры информатики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Рецензент: **Акимов Валерий Алексеевич,**
кандидат физико-математических наук, доцент кафедры математических методов в строительстве учреждения образования «Белорусский национальный технический университет»

Защита диссертации состоится «27» апреля 2022 г. года в $14^{\underline{00}}$ 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. 308-5, тел. 293-85-91, e-mail: inform@bsuir.by

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

КРАТКОЕ ВВЕДЕНИЕ

В настоящее время существует проблема наличия множества баз данных, содержащих неоднородную информацию. Как следствие возникают проблемы при попытках получить целевую информационную картину о предметной области, описанной в данных источниках. Одна из наиболее частых проблем, связанных с отсутствием Enterprise Data Warehouse, заключается в сложности создания отчетов пользователями. Так, часто те, кому нужна информация или требуются данные, которые они используют, должны ждать отчета, основанного на внешних источниках данных. В настоящее время руководители должны иметь возможность получить доступ к нужной им информации сразу же, как только она им понадобится.

Следовательно предприятия нуждаются в системах, способных осуществлять интеграцию и верификацию данных между различными системами. Наличие данного типа систем позволит упростить осуществление контроля над данными, и позволит пользователям получать наиболее актуальную информацию о предметной области в форме отчетов. В качестве такой системы отлично подойдет построение Data Warehouse. Стек технологий, использующихся с хранилищем данных, включает в себя средства и методы извлечения, преобразования и загрузки информации, а также компоненты Business Intelligence. Реализация Enterprise Data Warehouse, решает проблему интеграции нескольких систем в один общий источник информации с различными ролями, которые позволяют получить доступ к данным на разных уровнях.

Наиболее важной причиной для создания Enterprise Data Warehouse является улучшение системы отчетности. Объединяя данные из всех источников в Enterprise Data Warehouse, мы получаем универсальный источник для получения данных пользователями. Отчетность будет основываться на единой базе данных, а не на отдельных системах. Используя Enterprise Data Warehouse, человеку, создающему отчеты, не потребуется изучать несколько баз данных или пытаться объединить данные из нескольких баз данных, что иногда может являться довольно трудоемкой и долгой задачей. Помимо наличия единой базы данных для создания отчетов, существуют и другие причины не использовать данные из транзакционных баз данных, а именно: создание отчетов из транзакционной базы данных может замедлить общую производительность базы данных за счет использования ресурсов, уже выделенных для системы.

Конечной причиной разработки Enterprise Data Warehouse является то, что данные существуют для пользователей, поэтому согласованность является ключевым фактором использования данных для бизнеса. Независимо от источника данных, общая модель данных дает пользователям интерфейс для получения данных, которые им необходимы для достижения своих целей. Эта модель данных существует для удовлетворения потребности бизнеса в получении данных в форме отчетов для принятия более эффективных решений.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы является разработка программного модуля для хранения и обработки финансовых данных из различных источников для работы предприятия. Данная система позволит собрать данные из различных источников и предоставит унифицированный интерфейс, при помощи которого пользователи будут взаимодействовать с данными. Из которого можно было бы генерировать отчеты разного уровня, в соответствии с бизнес-требованиями, намного эффективнее, чем из источников данных.

Для достижения поставленной цели необходимо решить следующие задачи:

1 Проанализировать методы и средства построения хранилища данных, определить наиболее подходящий метод организации хранилища для текущего проекта.

2 Разработать архитектуру хранилища данных.

3 Реализовать процесс загрузки данных в хранилище.

4 Протестировать разработанную систему.

Объектом исследования являются системы хранения и процессы обработки финансовой информации.

Предметом исследования является инструментальные средства создания хранилищ данных, а также средства обработки данных с возможностью построения финансовой отчетности на их основе.

Основной гипотезой, положенной в основу диссертационной работы, является улучшение формирования финансовой отчетности за счет разработки хранилища данных. Хранилища данных способны осуществлять интеграцию и верификацию данных между различными системами. Их наличие позволит упростить осуществление контроля над данными, и позволит пользователям получать наиболее актуальную информацию о предметной области в форме отчетов.

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Разработано хранилище данных, а также процесс загрузки в него данных из различных источников. Произведено тестирование полученной системы.

Публикации результатов диссертации

По теме диссертации опубликовано 2 печатных работ, из них 1 статья в рецензируемом издании и 1 работа в сборнике трудов и материалов международных конференций.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, четырех глав, заключения, списка использованных источников, списка публикаций автора и приложения. В первой главе представлен анализ предметной области, разобраны основные методы построения хранилищ данных, выполнено их сравнение и определение наиболее подходящего для достижения цели данной диссертации. Вторая глава посвящена краткому обзору используемых при разработке приложения технологий. В рамках третьей главы разработана архитектура системы хранилища данных и реализован процесс ее заполнения. В четвертой главе приведены результаты тестирования разработанного программного средства.

Общий объем работы составляет 63 страниц, из которых основного текста – 45 страниц, 22 рисунка на 10 страницах, 9 таблиц на 5 страницах и список использованных источников из 36 наименований на 3 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе представлен анализ предметной области, разобраны основные методы построения хранилищ данных, выполнено их сравнение и определение наиболее подходящего для достижения цели данной диссертации.

Транзакционные системы не подходят для анализа больших объемов финансовой информации. В транзакционной системе информация актуальна только на момент обращения к базе данных, в следующий момент времени по тому же запросу Вы можете получить совершенно другой результат. Интерфейс транзакционных систем рассчитан на проведение строго определенных операций, и возможности получения результатов на нерегламентированный запрос сильно ограничены. Ответом на возникшую потребность стало появление новой технологии организации баз данных – технологии хранилищ данных (Data Warehouse).

Хранилище данных, Data Warehouse – предметно-ориентированная информационная база данных, специально разработанная и предназначенная для подготовки отчётов и бизнес-анализа с целью поддержки принятия решений в организации. Data Warehouse строится на базе систем управления базами данных и систем поддержки принятия решений. Данные, поступающие в хранилище данных, как правило, доступны только для чтения. Данные из OLTP- системы копируются в хранилище данных таким образом, чтобы построение отчётов и OLAP-анализ не использовал ресурсы транзакционной системы и не нарушал её стабильность. Как правило, данные загружаются в хранилище с определённой периодичностью, поэтому актуальность данных может несколько отставать от OLTP -системы.

Принципы организации хранилища:

– проблемно-предметная ориентация. Данные объединяются в категории и хранятся в соответствии с областями, которые они описывают, а не с приложениями, которые они используют;

– интегрированность. Данные объединены так, чтобы они удовлетворяли всем требованиям предприятия в целом, а не единственной функции бизнеса;

– некорректируемость. Данные в хранилище данных не создаются: т.е. поступают из внешних источников, не корректируются и не удаляются;

– зависимость от времени. Данные в хранилище точны и корректны только в том случае, когда они привязаны к некоторому промежутку или моменту времен

Концептуально модель хранилища данных можно представить в виде схемы, представленной на рисунке 1.

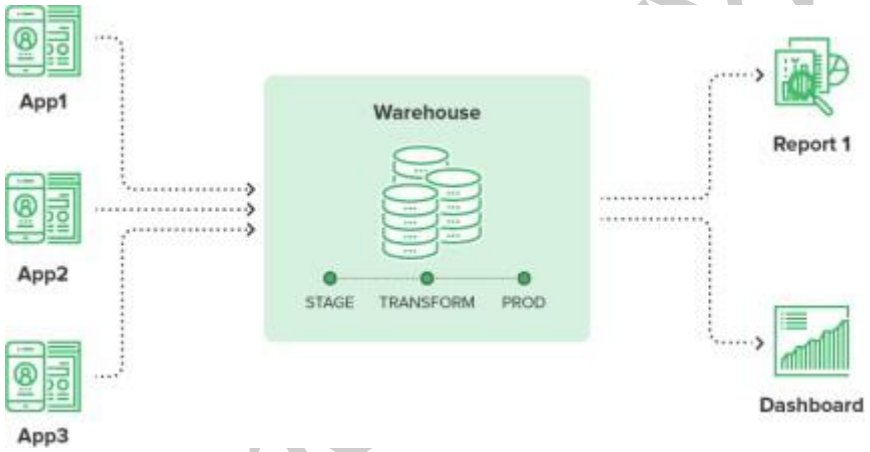

Рисунок 1 – Концептуальная архитектура системы

Хотя большинство баз данных, которые используются в среде, будут иметь свое приложение, расположенное поверх него для удобства, хранилище данных тоже можно рассматривать просто как очень большую базу данных, не имеющую традиционного уровня приложений. Тем не менее, есть приложения, которые являются частью хранилища данных, но они дополняют склад как компоненты, которые обеспечивают другую цель. Во-первых, имеется приложение ETL (Извлечение, преобразование и загрузка), которое является частью решения, которое перемещает данные из базы данных источников транзакций. Некоторые популярные приложения ETL – это Datastage, приложение OBIEE от Oracle и Informatica

Самыми известными и основными подходами к проектированию хранилища данных являются метод Инмона и метод Кимбалла. Оба подхода, как Инмона, так и Кимбалла, подходят для создания успешного Data Warehouse. Нельзя обобщать и говорить, что один подход лучше другого. У обоих есть свои преимущества и недостатки, и оба они прекрасно работают в разных сценариях. .Вот решающие факторы, которые могут помочь архитектору выбрать между ними:

– требования к отчетности. Если требования к отчетности являются стратегическими и общеорганизационными и требуется интегрированная отчетность, то Инмона подойдет лучше всего. Если же требования к отчетности носят тактический характер и ориентированы на бизнес-процесс / команду, то следует выбратьКимболла;

– срочность проекта. Если у организации достаточно времени, чтобы дождаться первого запуска хранилища данных (скажем, от 4 до 9 месяцев), то можно использовать подход Инмона. Если для запуска и запуска хранилища данных требуется очень мало времени (скажем, от 2 до 3 месяцев), то лучше всего использовать подход Кимбалла;

– будущий кадровый план. Если компания может позволить себе иметь большую команду специалистов для обслуживания хранилища данных, то можно использовать метод Инмона. Если будущий набор команды ограничен, то Кимболл больше подходит;

– частота изменений. Если ожидается, что требования к отчетности будут меняться быстрее, а исходные системы, как известно, нестабильны, то подход Инмона работает лучше, поскольку он более гибкий. Если требования и исходные системы относительно стабильны, можно использовать метод Кимбалла.

– организационная культура. Если спонсоры хранилища данных и менеджеры фирмы понимают ценность предложения хранилища данных и готовы принять долгосрочную выгоду от инвестиций в хранилище данных, то лучше подход Инмона. Если спонсоры не заботятся о концепциях, но хотят, чтобы ешение улучшало отчетность, тогда достаточно подхода Кимбалла

ETL (от англ. Extract, Transform, Load – дословно «извлечение, преобразование, загрузка») – комплекс методов, реализующих процесс переноса исходных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных. Целью ETL приложения является извлечение информации из одного или нескольких источников, преобразование ее в формат, поддерживаемый системой хранения и обработки, которая является получателем данных, а затем загружают в нее преобразованную информацию.

Независимо от особенностей построения и функционирования ETLсистема должна обеспечивать выполнение трех основных этапов процесса переноса данных (ETL-процесса): извлечение, преобразование и загрузка данных. Преобразованные данные записываются в соответствующую систему хранения.

Во второй главе приведен обзор используемых в данной работе технологий. Обоснован выбор данных инструментов, рассмотрены их достоинства и недостатки.

Для построения хранилища данных использовалась база данных Db2 on IBM Cloud. DB2 — это семейство систем управления реляционными базами данных, выпускаемых корпорацией IBM. В настоящее время СУБД DB2 представлена на нескольких платформах: Linux, Unix, z/OS, IBM Cloud.

Для построения ETL процесса был выбран IBM InfoSphere DataStage. DataStage является инструментом интеграции данных для проектирования, разработки и выполнения заданий, которые перемещают и преобразуют данные. Это ведущая платформа ETL, которая объединяет данные в нескольких корпоративных системах. Она использует высокопроизводительную параллельную инфраструктуру, доступную на месте или в облаке. Масштабируемая платформа обеспечивает расширенное управление метаданными и подключение к предприятиям.

DbVisualizer представляет собой многоплатформенный инструмент, предназначенный для администрирования реляционных БД ведущих производителей с использованием механизма доступа к данным JDBC. Среди поддерживаемых этим продуктом СУБД —SAP DB, Cache, MySQL, СУБД компаний IBM, Oracle, Microsoft, Pervasive, Sybase, а также ряд менее известных СУБД. DbVisualizer доступен для платформ Windows, Linux, Mac OS X и Solaris.

В третьей главе предложена разработанная архитектура хранилища данных. Основным источником данных служит операционная база данных приложения финансовой отчетности и планирования. Данные в этой операционной базе могут быть как введены пользователями через веб-интерфейс приложения, так и загружены из сторонних систем (баз данных филиалов компании). Эти сторонние системы также выступают источником данных и для разрабатываемого хранилища данных (рисунок 2)

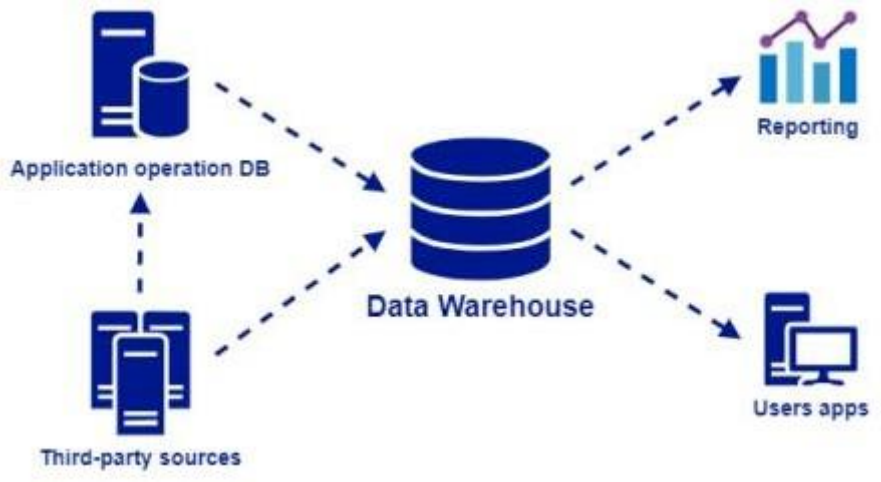

Рисунок 2 – Высокоуровневая архитектура системы

Для построения хранилища был выбран метод Кимбалла, во-первых, потому, что в данном проекте построение отчетности в первую очередь ориентированно на бизнес-процессы, а не на увеличение ее интегрированности. Во-вторых, потому что построение модели по Имболу требует большей команды высококвалифицированных профессионалов. И в-третьих, метод Кимбалла уменьшает время, необходимое на развертывание и настройку хранилища.

Как уже отмечалось в пункте, посвященному описанию хранилища по Кимбаллу, основой такого хранилища является звездообразная структура базы данных. Разработанную в данном проекте звездообразную схему хранилища данных можно увидеть ниже на рисунке 3.

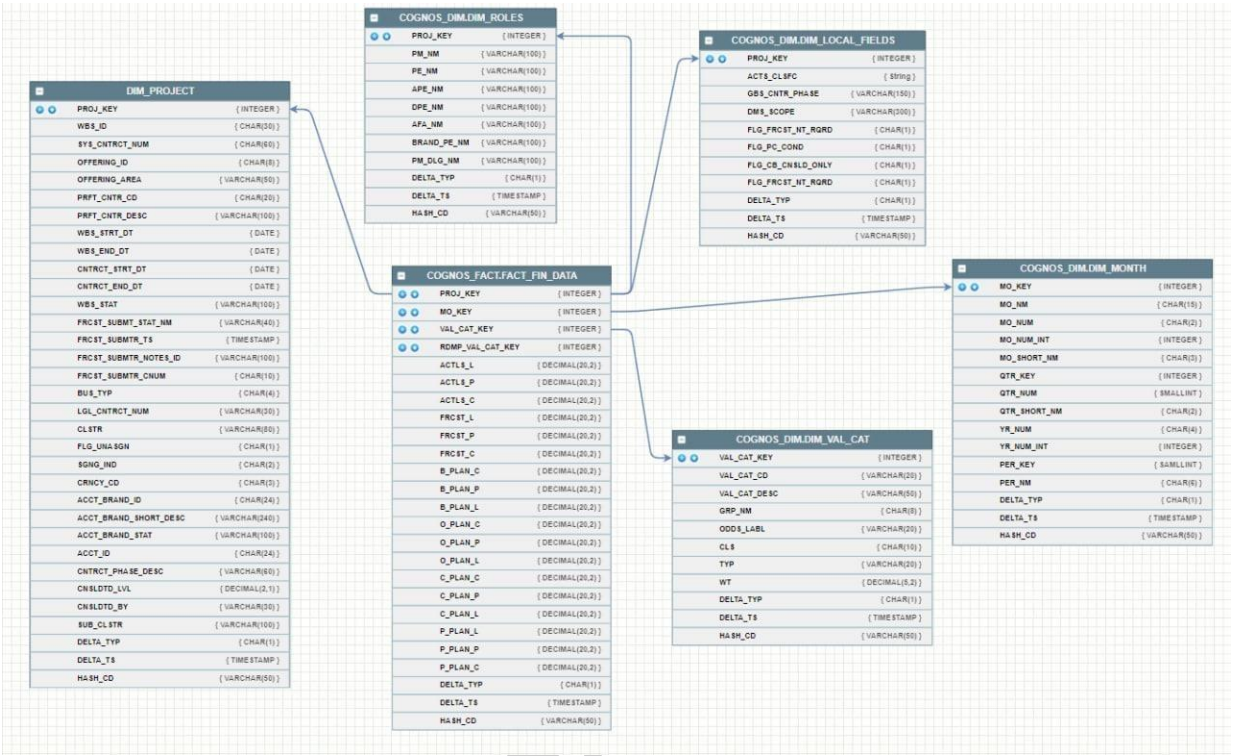

Рисунок 3 – Звездообразная схема хранилища данных

Для заполнения хранилища данных необходимо разработать ETL процесс, который обеспечит эффективный сбор информации из различных источников. Основными этапами разработки ETL процесса для данного хранилища является проектирование следующих его частей:

– загрузка данных в промежуточную область хранения;
– заполнение таблиц измерений;
– заполнение таблиц фактов.

В работе подробно рассмотрено создание всех этапов, на примере загрузки в систему данных, необходимых для «Forecast detailed» отчета. В пояснительной записке приведены структуры таблиц хранилища данных, а также Sql-код их заполнения.

Реализация заполнения данных таблиц разрабатывалась с помощью применения инструмента IBM DataStage. На рисунке 4 представлен пример джобы, выполняющей загрузку в таблицу WBS_ELEMENT. Данная джоба выгружает данные из источника, сравнивает с данными, что лежат в STG, и обновляет только изменившиеся данные. На рисунке показаны следующие DataStage элементы: 4 «DB2 connector», «check_sum», «change_capture», «transformer», «funnel».

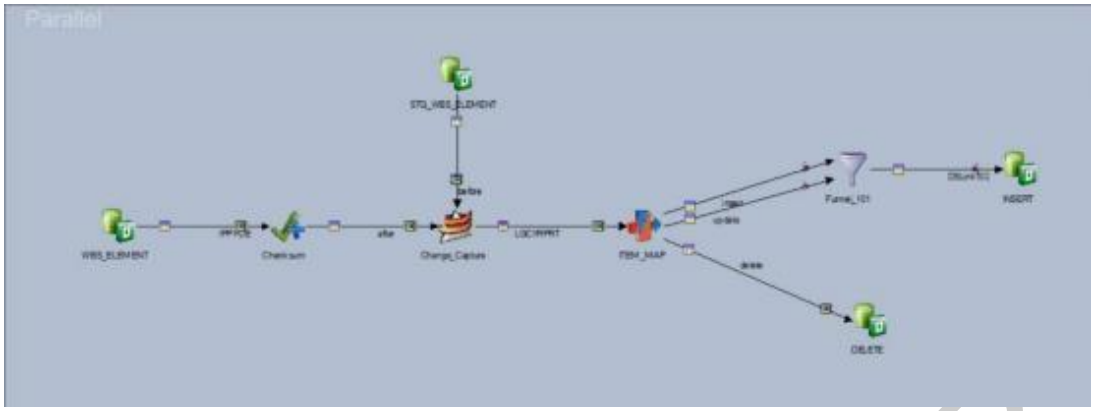

Рисунок 4 – Загрузка таблицы WBS_ELEMENT

В четвертой главе представлены результаты тестирования разработанного хранилища данных. Результаты проиллюстрированы на рисунке 5.

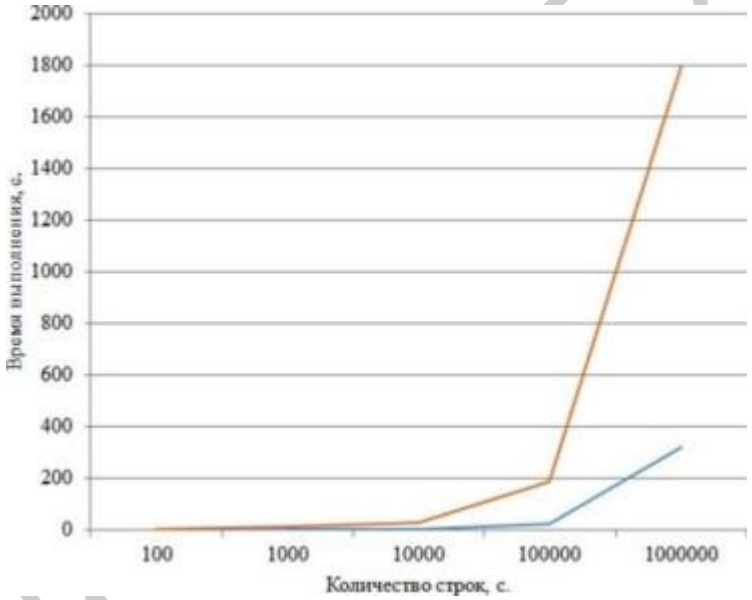

Рисунок 5 – Сравнение времени выполнения отчетов.

Из результатов видно, что время время выполнения отчетов существенно различается в источнике и Data Warehouse. При этом существенно выигрывает Data Warehouse. Исходя из этого можно сделать вывод, что использование Data Warehouse позволяет существенно уменьшить время выполнения отчетов.

ЗАКЛЮЧЕНИЕ

В результате работы над магистерской диссертацией были выполнены следующие задачи.

Рассмотрены основные принципы работы DataWarehouse и их взаимодействие с Business Intelligence. Обоснована необходимость в

использовании хранилищ данных для взаимодействия с большими объемами данных. Рассмотрены основные методы организаций DataWarehouse, их достоинства и недостатки.

Осуществлена реализация Data Warehouse в контексте управления проектами, для осуществления интеграции систем и улучшения качества отчетности. Данная система позволяет собрать данные из различных источников и предоставить унифицированный интерфейс, при помощи которого пользователи будут взаимодействовать с данными.

Разработан ETL-процесс для загрузки данных в созданное хранилище данных. Проиллюстрированы возможности программы IBM DataStage.

Произведено тестирование разработанного хранилища данных. Из полученного Data Warehouse можно генерировать отчеты разного уровня, в соответствии с бизнес-требованиями, намного эффективнее чем из источника данных.

Подводя итоги, можно резюмировать, что благодаря данной разработке существенно увеличилась скорость доступа к данным. Пользователи получили возможность оперировать данными из нескольких источников и на основании этих данных производить анализ, который раньше был им не доступен. И благодаря этому анализу принимать более правильные бизнес- решения.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

1. Гусаковская, Е. Г. Применение IBM InfoSphere DataStage в создании процесса обработки и хранения корпоративной информации / Е. Г. Гусаковская // Студенческий форум: электрон. научн. журн. 2020. № 1(94). Режим доступа: https://nauchforum.ru/journal/stud/94/64693.\

2. Козуб В. Н. Оценка масштабируемости облачных хранилищ данных. / В.Н. Козуб, З.И. Бессараб, Е.Г. Гусаковская // Big Data и Анализ Высокого Уровня. Сборник материалов VI международной научно-практической конференции. – Минск: БГУИР, 2020. – часть 3, с. 439 - 448

.