УДК 004.65

Л. В. РУДИКОВА

ОБ ОБЩЕЙ КОНЦЕПЦИИ УНИВЕРСАЛЬНОЙ СИСТЕМЫ ХРАНЕНИЯ И ОБРАБОТКИ ДАННЫХ ПРАКТИКО-ОРИЕНТИРОВАННОЙ НАПРАВЛЕННОСТИ

Гродненский государственный университет имени Янки Купалы

Развитие подходов и концепции к построению систем, связанных с накоплением данных в хранилище и последующим использованием алгоритмов Data Mining является особо перспективным в силу того, что белорусский сектор соответствующих ИТ-разработок находится еще на стадии формирования. В статье рассматривается общая концепции построения системы накопления и анализа данных практико-ориентированной направленности, основанная на технологии складирования данных. Основным аспектом в концепции проектирования универсальной системы на уровне хранения и работы с данными является подход с использованием расширяемого хранилища данных на основе универсальной платформы хранимых данных, который предоставляет доступ для хранения и последующего анализа данных различной структуры и различных предметных областей, имеющих точки (узлы) стыковки и расширенный функционал с возможность выбора структуры для хранения данных и последующей внутрисистемной интеграцией. Приводятся общая архитектура универсальной системы хранения и обработки данных указанной направленности, выделяются структурные составляющие. Основными компонентами архитектура универсальной системы для хранения и обработки данных практико-ориентированной направленности являются: оперативные источники данных; ETL-процесс; хранилище данных; подсистема анализа; пользователи. Важное место в структуре системы занимает аналитическая обработка данных, поиск информации, хранение документов, а также предоставление программного интерфейса для доступа к функциональности системы извне. Универсальная система на основе предлагаемой концепции позволит собирать достаточно обширные сведения по различным предметным областям, а также получать необходимые аналитические сводки, проводить обработку данных и применять соответствующие методы и алгоритмы Data Mining.

Ключевые слова: *универсальная система, данные практико-ориентированной направленности, технология складирования данных, общая архитектура, оперативные источники данных; ETL-процесс; хранилище данных; подсистема анализа*

Введение

В настоящее время развитие методов записи и хранения данных привели к огромному росту объемов накопленной, практически, необработанной информации. Конечно, при рассмотрении тех или иных аспектов предметных областей можно указать ресурсы и средства, которые используются для анализа накопленной информации. Однако, целый ряд направлений деятельности различных структур общества требуют построения концепции, а, в дальнейшем, разработки и использования соответствующих систем накопления, расширенного поиска и анализа больших объемов данных.

Развитие подходов и концепции к построению систем, связанных с накоплением данных в хранилище и последующим использованием алгоритмов Data Mining [1–3] является особо перспективным, т. к. белорусский сектор соответствующих ИТ-разработок находится еще на стадии формирования.

В силу вышеизложенного особый интерес представляет разработка объектной структурно-аналитической методологии, а также непосредственное создание и апробация программных комплексов, которые позволят накапливать данные в общее хранилище, осуществлять необходимый направленный поиск, а также проводить аналитические исследования, включая интеллектуальный анализ данных.

Рассматриваемая тематика, связанная с разработкой общей концепции построения систем накопления и анализа данных [4], включает

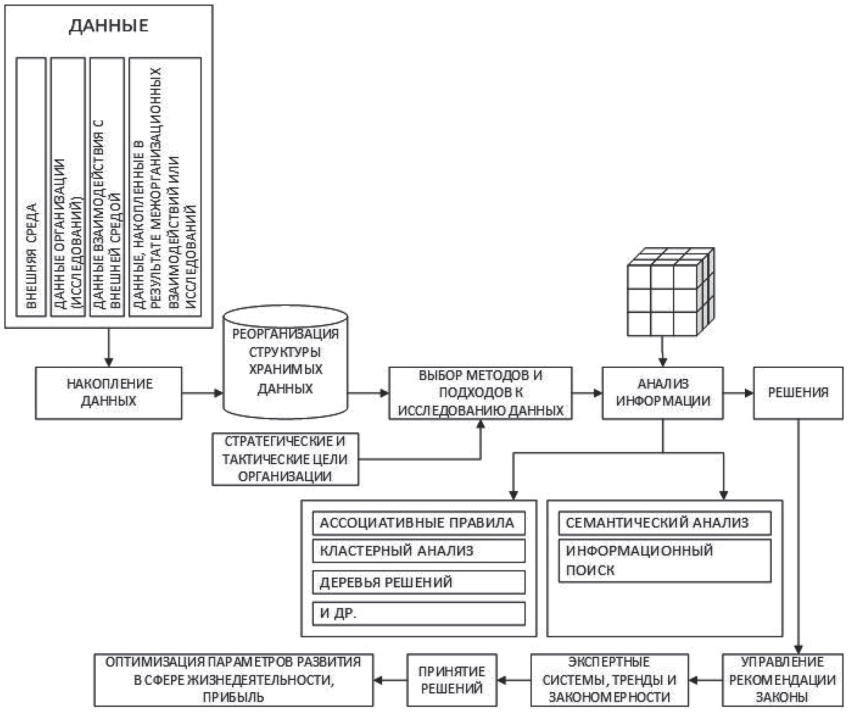

Рис. 1. Упрощенная схема функционирования организации

следующие направления исследований. Во-первых, исследование предметной области, связанной с объектами художественной ценности и соответствующей лазерной экспрессной экспертизой: анализ полученных результатов по структурированию информации, построение модели для хранилища данных, возможные варианты обработки информации. Во-вторых, сбор и накопление информации, касающейся персональных данных людей, прежде всего, известных личностей (в дальнейшем предполагается провести исследования в аспекте востребованности специалистов в той или иной области, кадрового потенциала выпускников и тенденций, связанных с подготовкой высококвалифицированных специалистов по конкретным специальностям). И, в-третьих, направление, характеризующее некоторые аспекты предметной области демографических и миграционных данных [5, 6].

Общая концепция хранилища данных для хранения и анализа данных практико-ориентированной направленности

На рис. 1 показана упрощенная схема функционирования организации (с учетом современных тенденций к интеграции данную схему можно трактовать также и как основу единой структуры в рамках определенной федеральной или государственной структуры) на базе накопленной информации (научной, производственной, социальной и т. п.) и место анализа непрерывно поступающей информации.

Отметим, что на текущий момент предложены общие подходы к созданию универсальной системы, которая позволяет поддерживать различные этапы, связанные с проведением лазерной экспрессной экспертизы, автоматизировать процессы хранения и поиска данных с целью их дальнейшей обработки и получения требуемых экспертных заключений [7–12]. Проводятся работы по комплексному исследованию предметных областей, расширению структурно-аналитической методологии построения подобного рода программных систем, результаты которых можно применить и к более общим программным комплексам, предназначенным для обработки статистических данных большого объема практико-ориентированной направленности.

Основная концепция предлагаемой универсальной расширяемой системы для хранения и анализа данных практико-ориентированной направленности основана на технологии складирования данных. Разработка хранилища данных предполагается с учетом того, что в конечном итоге универсальная система будет

предоставлять большой комплекс услуг соответствующим группам пользователей. Очевидно, что ресурсоемкость системы будет расти по мере того как система будет наполняться данными и обслуживать все большее количество пользователей. Немаловажную роль в плане определения концепции построения системы также играет необходимость осуществления аналитической обработки поступающих данных, поиска информации, хранение документов, а также предоставление программного интерфейса для доступа к функциональности системы извне [13, 14].

Основным аспектом в концепции проектирования универсальной системы на уровне хранения и работы с данными является подход с использованием хранилища данных (Data Warehouse) [15] – предметно-ориентированной информационной базы данных, построенной на основе схемы «созвездие фактов», специально разработанной и предназначенной для подготовки отчётов и бизнес-анализа с целью поддержки принятия решений по различным направлениям указанных тематик. Данные, которые поступают в хранилище данных, доступны, в основном, только для чтения. Из OLTP-системы необходимая информация копируется в хранилище данных таким образом, чтобы итоговые построенные отчеты и OLAP-анализ не обращался к ресурсам транзакционной системы и, таким образом, не нарушал её стабильность. Предполагается, что данные загружаются в хранилище с периодичностью в неделю или декаду, поэтому актуальность данных может несколько отставать от OLTP-системы. Ниже предлагается обобщенная архитектура для универсальной системы хранения и анализа данных на базе расширяемого хранилища данных. В данном случае, в качестве *расширяемого* хранилища данных предлагается подход на основе универсальной платформы хранимых данных, который предоставляет доступ для хранения и последующего анализа данных различной структуры и различных предметных областей, имеющих точки (узлы) стыковки и расширенный функционал с возможность выбора структуры для хранения данных и последующей внутрисистемной интеграцией. Отметим также, что предполагается построение некоторой универсальной системы, которая при правильном ее проектировании и выборе соответствующих методов и подходов к обработке данных может разрастись до глобального применения за счет гибкости в плане расширения ее функциональности, обилию предоставляемых аналитических данных, максимально универсальным структурам хранения данных, устойчивости к большим нагрузкам и т. д. Кроме непосредственно подсистемы хранения и обработки поступающих со стороны ее пользователей данных, в системе будет располагаться большое количество вспомогательной информации, необходимой для поддержки работы подсистем, выполняющих структурный и экспертный анализ. Такого рода информация также будет обособлена и вынесена в отдельную базу данных, структурированную для максимально быстрого поиска информации и доступа к ней.

Общая архитектура универсальной системы хранения и обработки данных практико-ориентированной направленности

На рис. 2 представлена разработанная архитектура для универсальной системы хранения и обработки данных практико-ориентированной направленности.

Основными компонентами архитектуры универсальной системы для хранения и обработки данных практико-ориентированной направленности являются: оперативные источники данных; ETL-процесс; хранилище данных; подсистема анализа; пользователи.

Охарактеризуем основные компоненты системы.

Оперативные источники данных включают различные документы и данные, которые обрабатываются OLTP-системами.

ETL-процесс представляет собой процесс извлечения (получения) информации из OLTP-систем (баз данных), затем ее дальнейшее преобразование к формату хранимых данных в хранилище, и непосредственной загрузки данных в хранилище данных. Для поддержки ETL-процесса используются соответствующие программы, которые позволяют извлекать данные из исходной базы данных, преобразовывать их в соответствии с требованиями и загружать в хранилище. Отметим, что для извлечения данных из исходной базы данных можно использовать как готовое программное обеспече-

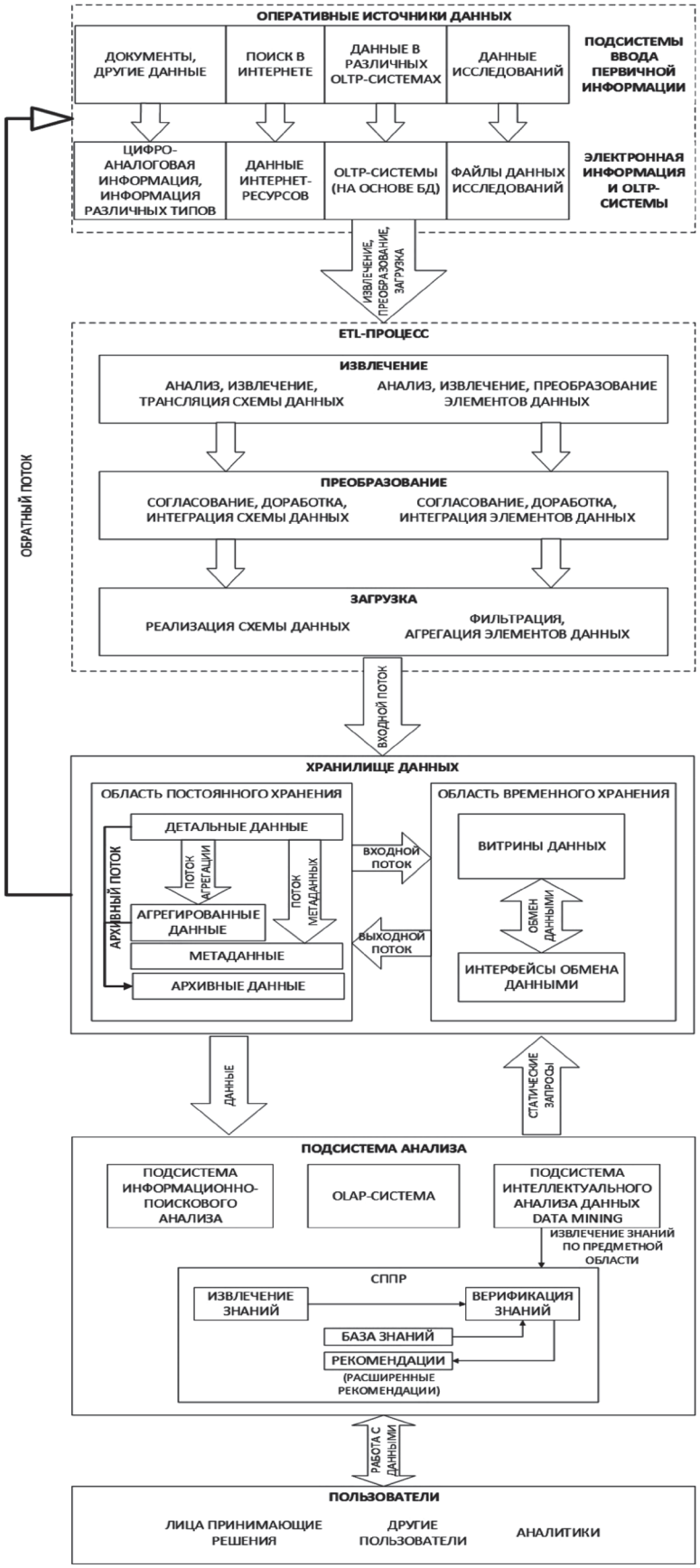

Рис. 2. Общая архитектура универсальной системы для хранения и обработки данных практико-ориентированной направленности

ние, так и разрабатывать собственные решения с учетом требований итоговых форматов данных.

Архитектура области хранения данных может проектироваться в зависимости от требований представления и использования итоговой информации. В рассматриваемом случае архитектуру *корпоративного хранилища данных* целесообразно проектировать с учетом следующих областей.

1) Область временного хранения данных (Staging Area) – будет использоваться для временного хранения данных, которые извлекаются из транзакционных систем; таким образом, эта область – промежуточный слой между OLTP-системами и хранилищем данных, которая состоит из:

– витрин (киосков) данных (Data Marts) – необходимые представления данных для их анализа конечными пользователями (в соответствии с требуемыми схемами данных, например, типа «снежинка», «звезда», «созвездие фактов»); витрины данных предназначены для поддержки итоговых целей пользователей, компаний и приложений;

– интерфейсов обмена данными с другими системами и приложениями (Data Exchange Interface или Feedback Area), представляющих собой специальные таблицы, в которых хранятся подготовленные для передачи в другие приложения компании или организации данные из области постоянного хранения данных.

2) Область постоянного хранения данных состоит из:

– детальных данных (System of records), т. е. области хранения детальных данных хранилища в соответствии со структурой модели данных хранилища;

– агрегированных данных (Summary area), т. е. сгруппированных по определенным показателям детальных данных;

– метаданных (Metadata) – данных, которые описывают структуры для хранения данных, т. е. это стандартные описания структур таблиц, взаимосвязей, ограничений, правил секционирования, описание витрин, данных и т. п.

– архивных данных (Archive data) – данных, которые не используются в активных отчетах, исследованиях и обработках, но которые при необходимости могут быть извлечены из хранилища; как правило, будут заданы определенные критерии для помещения данных хранилища в архивные таблицы.

Подсистема анализа также важная часть предлагаемого архитектурного решения, основными требования к которой являются расширяемость и модульность структуры.

Основные программно-системные модули, которые добавляются в подсистему анализа, зависят от конкретных задач пользователей и аналитических систем. Как правило, к основным модулям анализа данных можно отнести подсистему информационно-поискового анализа, OLAP-систему, подсистему интеллектуального анализа данных Data Mining, систему поддержки принятия решений.

В *подсистему информационно-поискового анализа* необходимо включить информационный поиск, визуальный поиск и семантический поиск. постановка четких запросов к базе данных и получение соответствующей информации – это главные задачи, решаемые подсистемой информационного и визуального поиска.

Использование технологии *OLAP* в предлагаемой системе – это соответствующий подбор программных решений класса Business Intelligence для быстрого создания и настройки многомерных кубов с целью получения итоговых агрегированных запросов по конкретному направлению, соответствующих направлению предметной области.

Подсистема интеллектуального анализа данных Data Mining предполагает возможность с использованием соответствующих методов обработки данных обнаружения неизвестных ранее нетривиальных и практически полезных знания, которые можно интерпретировать соответствующим образом и на этой основе принимать соответствующие решения по направлениям деятельности, указанным выше. Как правило, Data Mining использует следующие методы: нейронные сети, деревья решений, алгоритмы кластеризации, в том числе и масштабируемые, алгоритмы обнаружения ассоциативных связей между событиями и т. д.

Основное назначение *системы поддержки принятия решений* (СППР) состоит в получении рекомендаций на основе изучения имеющихся исторических и текущих данных о состоянии требуемого объекта исследования и сравнение этой информации с информацией, которая хранится в базе данных системы. Ос-

новные функции СППР – это извлечение знаний, их верификация и формирование рекомендаций требуемого образца. В СППР обязательно должна присутствовать возможность обновления знаний с учетом текущего момента времени (актуализация знаний).

К системе могут иметь доступ различные группы пользователей, которым доступен и различный функционал системы. Отметим, что, кроме администраторов системы, можно указать также следующие группы пользователей: лица, принимающие решения, аналитики и другие пользователи.

Лица, принимающие решения (ЛПР) – это группа пользователей, от решения которых зависит анализ объективной составляющей ситуации или явления; выявление предпочтений ЛПР, генерация возможных решений, оценка возможных альтернатив, анализ последствий принимаемых решений, выбор лучшего варианта и т. д.

Аналитики – пользователи, которым необходимо сформулировать и проверить конкретные гипотезы, а также создавать требуемые запросы различного плана и подготовить отчеты по конкретным направлениям предметной области.

Другим пользователям, как правило, доступна общая информация, связанная с универсальной системой и для них ограничены возможности по обработке имеющихся данных предлагаемой системы.

Однако, стоит помнить, что, в зависимости от целей использования системы, всегда можно определить конкретного пользователя и предоставить ему определенные права на доступ и обработку информации, а также – к отдельным модулям и функционалу системы.

Рекомендации по выбору программных средств

Разрабатываемая система должна включать в себя единое централизованное хранилище данных, систему обработки данных и систему отчетности. В хранилище данных информация хранится в специально реорганизованном виде в соответствии с разработанной структурой хранения, содержащей необходимые измерения и агрегированные факты предметной области. Для реализации и поддержки хранилища данных рекомендуется использовать MS SQL Server. MS SQL Server используется в системах, осуществляющих хранение больших объемов, данных с требуемой политикой безопасности. MS SQL Server имеет средства аналитической обработки многомерных моделей, данных (OLAP) и сбора релевантной информации, которые входят в состав Microsoft Analysis Services. В SQL Server имеется поддержка. NET Framework, в силу чего хранимые процедуры баз, данных можно реализовывать на любом языке платформы. NET, используя полный набор библиотек, которые доступны для. NET Framework.

В качестве средства для работы с OLAP и интеллектуальным анализом данных хорошим решением являются службы Microsoft SSAS.

Заключение

Разработка общей концепции и реализация универсальной Интернет-системы складирования и обработки данных практико-ориентированной направленности, связанных с различными видами деятельности людей, может быть рассмотрена в аспекте создания некоторого федерального хранилища данных, что, несомненно, является актуальной темой исследования. Интересным также представляется распространение методов и технологий бизнес-аналитики для научных исследований и получение соответствующих результатов, которые позволяют определять перспективы использования определенных ресурсов, возможные закономерности по имеющимся массивам данных, а также влияние определенных параметров на развитие социальных, исторических и экологических процессов. Система такого рода позволит собрать достаточно обширные сведения по различным предметным областям, а также получать необходимые аналитические сводки, проводить обработку данных и применять соответствующие методы и алгоритмы Data Mining.

Литература

1. **Devlin, B. A.** An Architecture for a Business and Information System / B. A. Devlin, P. T. Murphy. – IBM Systems Journal, 1988. – Vol 17, No 1. – P. 60–80.

2. **Inmon, W. H.** Building the Data Warehouse / W. H. Inmon // Third Edition. – John Wiley & Sons, Inc. New York, 2002. – 428 p.

3. **Kimbell, R.** The Data Warehouse Toolkit: The Complete Guide to Dimensional Data Warehouses / R. Kimbell, M. Ross // Second Edition. – J. Willey & Sons, 2002. – 447 p.

4. **«Примеры** реализации хранилищ данных предприятия» [Электронный ресурс] / Интернет-технологии. – Режим доступа: http://www.internet-technologies. ru/articles/article_994.html. – Дата доступа: 26.12.2016.

5. **Belyi, A.** Global multi-layer network of human mobility // Alexander Belyi, Iva Bojic, Stanislav Sobolevsky, Izabela Sitko, Bartosz Hawelka, Lada Rudikova, Alexander Kurbatski, Carlo Ratti / International Journal of Geographical Information Science. – Mode of access: [http://www.tandfonline.com/doi/full/10.1080/13658816.2017.1301455]. – Date of access: [14.04.2017].

6. **Белый, А. Б.** Данные сервиса Flickr и структура сообществ стран // А. Б. Белый, Л. В. Рудикова, С. Л. Соболевский, А. Н. Курбацкий / Международный конгресс по информатике, информационные системы и технологии = International Congress on Computer Sciens : Information Systems and Technologies : материалы междунар. науч. конгресса, Республика Беларусь, Минск, 24 окт. – 27 нояб. 2016 г. : редкол.: С. В. Абламейко (отв. ред.) [и др.]. – Минск : БГУ, 2016. – С. 851–855.

7. **Рудикова, Л. В.** О разработке системы для поддержки лазерной экспрессной экспертизы. Монография / Л. В. Рудикова. – LAP LAMBERT Academic Publishing, 2014. – 134 с.

8. **Рудикова, Л. В.** Особенности архитектурной реализации системы визуализации и обработки результатов спектрального анализа // Л. В. Рудикова / Доклады БГУИР. – Мн.: БГУИР, 2015. – № 1 (87) – С. 47–53.

9. **Рудикова, Л. В**. О разработке универсальной системы обработки данных, связанных с лазерной экспрессной экспертизой // Л. В. Рудикова / Системный анализ и прикладная информатика. – Мн.: БНТУ, 2015. – № 2. – С. 58–64.

10. **Рудикова, Л. В.** Формирование экспертных заключений с использованием лазерного метода спектрального анализа и специализированного программного обеспечения / Л. В. Рудикова, Е. В. Жавнерко, Н. Н. Курьян, Д. В. Лазарь // Доклады Белорусского государственного университета информатики и радиоэлектроники. – Мн.: БГУИР, 2016. – № 2. – С. 56–62.

11. **Рудикова, Л. В.** О проектировании системы для поддержки экспертизы объектов художественной ценности // Л. В. Рудикова / Информационные системы и технологии: управление и безопасность: сб. ст. IV международной заочной научно-практической конференции / Поволжский гос. ун-т сервиса. – Тольятти: Изд-во ПВГУС, 2016. – С. 154–167.

12. **Рудикова, Л. В.** О концепции универсальной системы хранения и обработки данных произведений художественной ценности // Л. В. Рудикова / Фундаментальные проблемы естествознания и техники. Серия: Проблемы исследования Вселенной. – Т. 37. № 2. – Санкт-Петербург, 2016. – С. 209–227.

13. **Барсегян, А. А.** Методы и модели анализа данных: OLAP и Data Mining / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод – СПб.: БХВ-Петербург, 2009. – 336 с.: ил.

14. **Паклин, Н. Б.** Бизнес-аналитика: от данных к знаниям / Н. Б. Паклин, В. И. Орешков – СПб.: Питер, 2009 год. – 624 с.

15. **Wrembel, R**. Data warehouses and OLAP: concepts, architectures, and solutions / R. Wrembel, C. Koncilia. – IRM Press, 2007. – P. 1–26.

References

1. **Devlin, B. A.** An Architecture for a Business and Information System / B. A. Devlin, P. T. Murphy. – IBM Systems Journal, 1988. – Vol 17, No 1. – P. 60–80.

2. **Inmon, W. H.** Building the Data Warehouse / W. H. Inmon // Third Edition. – John Wiley & Sons, Inc. New York, 2002. – 428 p.

3. **Kimbell, R.** The Data Warehouse Toolkit: The Complete Guide to Dimensional Data Warehouses / R. Kimbell, M. Ross // Second Edition. – J. Willey & Sons, 2002. – 447 p.

4. »**Implemented** data warehouses of company examples» [Electronic source] / Internet-technologies. – Mode of access: http:// www. internet-technologies. ru/articles/article_994. html. – Date of access: 26.12.2016.

5. **Belyi, A.** Global multi-layer network of human mobility // Alexander Belyi, Iva Bojic, Stanislav Sobolevsky, Izabela Sitko, Bartosz Hawelka, Lada Rudikova, Alexander Kurbatski, Carlo Ratti / International Journal of Geographical Information Science – Mode of access: [http:// www. tandfonline. com/doi/full/10.1080/13658816.2017.1301455]. – Date of access: [14.04.2017].

6. **Belyi, A. B.** Flickr service data and community structure of countries // A. B. Belyi, L. V. Rudikova, S. L. Sobolevsky, A. N. Kurbatski / International Congress on Computer Sciens : Information Systems and Technologies : materials of International scientific Congress, Republic of Belarus, Minsk, 24 October. – 27 Nov. 2016: rare: S. V. Ablameiko (editorial editors) [and others]. – Minsk: BSU, 2016. – P. 851–855.

7. **Rudikova, L**. About laser express expertise system implementation. Monography / Lada Rudikova. – LAP LAMBERT Academic Publishing, 2014. – 134 p.

8. **Rudikova, L.** Architecture's implementation features system of visualization and spectral analysis results processing // Lada Rudikova / Doklady BGUIR. – Minsk: BSUIR, 2015. – № 1 (87) – P. 47–53.

9. **Rudikova, L.** About universal system of laser express expertise data processing // Lada Rudikova / System analysis and applied information science. – Minsk: BNTU, 2015. – № 2. – P. 58–64.

10. **Rudikova, L. V.** Formation of expert conclusions using the laser method of spectral analysis and specialized software / L. V. Rudikova, E. V. Zhavnerko, N. N. Kuryan, D. V. Lazar // Reports of the Belarusian State University of Informatics and Radioelectronics. – Minsk: BSUIR, 2016. – № 2. – P. 56–62.

11. **Rudikova, L. V.** On the design of the system for the support of the examination of artwork objects // L. V. Rudikova / Information Systems and Technologies: Management and Security: Sat. Art. IV international correspondence scientific-practical conference / Povolzhsky State University of Service. – Togliatti: Publishing house of VGUS, 2016. – P. 154–167.

12. **Rudikova, L. V**. On the concept of an universal system for storing and processing artwork objects data // L. V. Rudikova / Fundamental Problems of Natural Science and Technology. Series: Problems of the study of the universe. – P. 37. № 2. – St. Petersburg, 2016. – P. 209–227.

13. **Barseghyan, A.** Methods and analysis data models: OLAP and Data Mining / A. Barseghyan, M. Kupriyanov, V. Stepanenko, I. Kholod – StP.: BHV-Petersburg, 2009. – 336 p.: il.

14. **Paklin, N.** Business- analytics: from data to knowledge / N. Paklin, V. Oreshkov. – StP.: Piter, 2009. – 624 p.

15**. Wrembel, R**. Data warehouses and OLAP: concepts, architectures, and solutions / R. Wrembel, C. Koncilia. – IRM Press, 2007. – P. 1–26.

Поступила	*После доработки*	*Принята к печати*
10.04.2017	*18.05.2017*	*10.06.2017*

Rudikova, L. V.

ABOUT THE GENERAL CONCEPT OF THE UNIVERSAL STORAGE SYSTEM AND PRACTICE-ORIENTED DATA PROCESSING

Approaches evolution and concept of data accumulation in warehouse and subsequent Data Mining use is perspective due to the fact that, Belarusian segment of the same IT-developments is organizing. The article describes the general concept for creation a system of storage and practice-oriented data analysis, based on the data warehousing technology. The main aspect in universal system design on storage layer and working with data is approach uses extended data warehouse, based on universal platform of stored data, which grants access to storage and subsequent data analysis different structure and subject domains have compound's points (nodes) and extended functional with data structure choice option for data storage and subsequent intrasystem integration. Describe the universal system general architecture of storage and analysis practice-oriented data, structural elements. Main components of universal system for storage and processing practice-oriented data are: online data sources, ETL-process, data warehouse, subsystem of analysis, users. An important place in the system is analytical processing of data, information search, document's storage and providing a software interface for accessing the functionality of the system from the outside. An universal system based on describing concept will allow collection information of different subject domains, get analytical summaries, do data processing and apply appropriate Data Mining methods and algorithms.

Keywords: *universal system, practice-oriented data, warehousing technology, common architecture, online data source, ETL-process, data warehouse, analysis subsystem*

Lada Rudikova is the Head of Modern Programming Technologies Department of Yanka Kupala State University of Grodno (YKSUG). Ph. D. degree in physical and math.

The main line of her scientific researches – management theory, information systems design, databases, CASE, data mining, business intelligence. She actively participates in international conferences. She is the author of more than 280 scientific works and books related to computer technology and data processing, a technical writer of the publishing house «BHV-St Petersburg».

Результаты работы получены в процессе выполнения ГПНИ «Разработка методологии и средств построения универсальных систем хранения, обработки и анализа структурированных данных большого объема практико-ориентированной направленности».