Сафрончик М.И.

ПРОЕКТИРОВАНИЕ И РЕАЛИЗАЦИЯ РЕШЕНИЙ ПО БИЗНЕС-АНАЛИТИКЕ НА БАЗЕ Microsoft SQL Server 2012

Учебное пособие

В четырех частях

Часть 1. Проектирование хранилищ данных

2015

Сафрончик М.И. (Саратовский Государственный Университет, факультет компьютерных наук и информационных технологий, кафедра математической кибернетики и компьютерных наук)

Проектирование и реализация решений по бизнес-аналитике на базе Microsoft SQL Server 2012: Учебное пособие в 4 ч. Ч.1. – LULU Inc., USA, 2015. – 124 с.

ISBN 978-1-329-78013-2

Данное учебное пособие посвящено изучению современных технологий создания решений для аналитической обработки данных и построению корпоративных информационных систем поддерживающих эти технологии. Рассмотрены примеры практической реализации. Приведены задания для самостоятельного выполнения.

Пособие может быть рекомендовано для студентов всех специальностей и направлений, связанных с информационными технологиями.

Рецензенты
доктор физико-математических наук, профессор **Д. К. Андрейченко**
кандидат физико-математических наук **И. А. Батраева**

Работа издана в авторской редакции

ISBN 978-1-329-78013-2 © Сафрончик М.И., 2015

Оглавление

Введение

В последние годы интерес к системам бизнес-аналитики постоянно растет. Ведь качество информационной поддержки деятельности руководителей и аналитиков является одним из важнейших факторов достижения успеха любого предприятия. Поэтому изучение современных методов анализа данных и построения корпоративных информационных систем, поддерживающих эти технологии, является весьма актуальной задачей.

В этом учебном пособии рассматриваются современные технологии создания решений для аналитической обработки данных и основные инструменты, используемые при построении таких систем, которыми располагает одна из самых популярных на сегодняшний день СУБД – SQL Server. Системы подобного класса позволяют строить аналитические отчеты, извлекая информацию из различных источников, и на основе этих отчетов принимать управленческие решения.

Цель данного учебного пособия – познакомить с основными направлениями аналитической обработки данных, методами построения решений для систем бизнес-анализа и механизмами SQL Server, поддерживающими эти решения.

Пособие предназначено для студентов всех специальностей и направлений, связанных с информационными технологиями.

В рамках данного учебного курса обсуждаются следующие вопросы:

- основы проектирования, реализации и обслуживания хранилищ данных и основные процессы, связанные с этими задачами;
- оперативная аналитическая обработка данных (от анг. online analytical processing – OLAP) и интеллектуальный анализ данных на основе хранилищ данных;
- различные средства построения аналитических отчетов.

Пособие состоит из четырех частей. Первая часть посвящена логическому и физическому проектированию хранилищ данных.

Весь материал разбит на главы, каждая из которых содержит теоретический материал и подробные примеры практической реализации с использованием соответствующего инструмента SQL Server. На протяжении всех глав описывается поэтапное создание образца хранилища данных. Структура этого хранилища несколько проще для восприятия, чем структура образца хранилища AdventureWorksDW, предоставляемого компанией Microsoft.

Для лучшего усвоения изученного материала и приобретения практических навыков создания хранилищ данных с помощью инструментов SQL Server в конце каждой главы приводятся задания для выполнения лабораторных работ, после выполнения которых, студенты научатся разрабатывать подходящую модель данных для хранилища и оптимизировать физическую модель хранилища.

В рамках данного лабораторного практикума студенты выбирают определенную предметную область и, выполняя последовательно все задания, в результате создают собственный готовый продукт – хранилище данных.

Методические указания для обучающихся

Изучение материалов данного курса предполагает наличие у студентов знаний и навыков работы с реляционными базами данных, включая разработку нормализованной модели данных, работу с компонентой Database Engine SQL Server 2012, создание таблиц и отношений между ними, а также некоторого знания языка T-SQL (диалект языка запросов SQL, используемый в SQL Server).

Для выполнения лабораторных работ на компьютере должен быть установлен экземпляр SQL Server 2012 с входом в систему, имеющим разрешения на создание новых баз данных, предпочтительно участником роли sysadmin. Для работы можно использовать практически любую локальную редакцию SQL Server (Standard, Enterprise, Business Intelligence и Developer), как 32-разрядную, так и 64-разрядную, но все преимущества инструментов бизнес-аналитики сосредоточены в редакциях Enterprise и Developer. Описание возможностей, поддерживаемых различными выпусками SQL Server 2012 можно найти на странице

https://technet.microsoft.com/library/cc645993(SQL.110).aspx. Если у вас нет доступа к имеющемуся экземпляру SQL Server, можно воспользоваться ознакомительной версией SQL Server 2012, которая действительна 180 дней. Загрузить ознакомительную версию можно со страницы https://technet.microsoft.com/ru-ru/evalcenter/hh225126. Минимальные требования к оборудованию и операционной системе для SQL Server 2012 представлены на странице http://msdn.microsoft.com/ru-ru/library/ms143506(v=sql.110).aspx.

Предисловие

Современные концепции построения аналитических решений

Основная цель построения любой информационной системы – это обеспечение эффективного управления предприятием. Аналитические системы позволяют управляющему звену проводить текущую оценку состояния предметной области, сравнивая внешние данные о состоянии рынка с данными, накопленными внутри компании, формулировать и описывать цели и задачи, определять методы и пути их достижения.

Такие решения принято называть системами бизнес-аналитики (от анг. Business Intelligence – BI). Они должны включать в себя средства сбора и обработки больших объемов данных и предоставлять пользователям широкий набор инструментов для анализа и извлечения знаний из этих данных в удобном для них виде.

Таким образом, процесс превращения данных в знания для поддержки принятия решений обеспечивается методами и средствами Business Intelligence.

Как правило, BI-системы включают в себя средства для решения четырех основных задач: хранения данных, интеграции данных, анализа данных и представления данных в удобном для аналитика виде.

Данные для бизнес-аналитики собираются из различных источников и организуются в структурированные специальным образом хранилища (анг. Data Warehouse – DWH) для более эффективной обработки аналитических запросов. Хранилища данных (ХД) содержат огромные объемы информации, охватывающие все сферы деятельности предприятия.

Для поддержки хранилищ данных в актуальном и согласованном виде и реализации некоторых дополнительных инструментов извлечения данных из разрозненных источников современные BI-системы содержат развитые средства интеграции данных.

Задачи анализа данных можно разделить на три класса по уровню извлекаемых знаний и используемыми аналитическими инструментами.

- ✓ Информационно-поисковые – применяется язык запросов к базе данных, активно обрабатывающей транзакции, позволяют извлечь лишь поверхностный слой знаний. В BI-системах в этом случае выполняются заранее определенные программистами запросы.
- ✓ Оперативно-аналитические (от анг. online analytical processing – OLAP) – используются средства оперативной аналитической обработки данных, что дает возможность вскрыть более глубокие пласты представленных в хранилище данных. В этом случае производится группировка и обобщение данных в любом виде, необходимом аналитику, и появляется возможность выполнять нерегламентированные заранее запросы (Ad-hoc query).
- ✓ Задачи интеллектуального анализа данных (Data Mining). Осуществляется поиск функциональных и логических закономерностей в накопленных данных, строятся модели и правила, которые объясняют найденные закономерности и/или (с определенной вероятностью) прогнозируют развитие некоторых процессов. Эта технология открывает поистине неисчерпаемые возможности для поиска самых глубинных пластов знаний и выявления скрытых закономерностей, что позволяет предприятиям оптимизировать свой бизнес за счет обнаружения неявных взаимозависимостей, предсказания важных маркетинговых факторов и прогноза стратегии ведения бизнеса [1].

Также в состав инструментов BI обязательно входят средства для создания различных видов отчетов (Reporting) и их рассылки.

Таким образом, системы Business Intelligence – это целый комплекс средств, позволяющих извлекать полезные знания из данных, собранных из разнородных источников, и представлять информацию в виде, удобном для бизнес-аналитики. На рис. 1 показана обобщенная архитектура системы бизнес-аналитики.

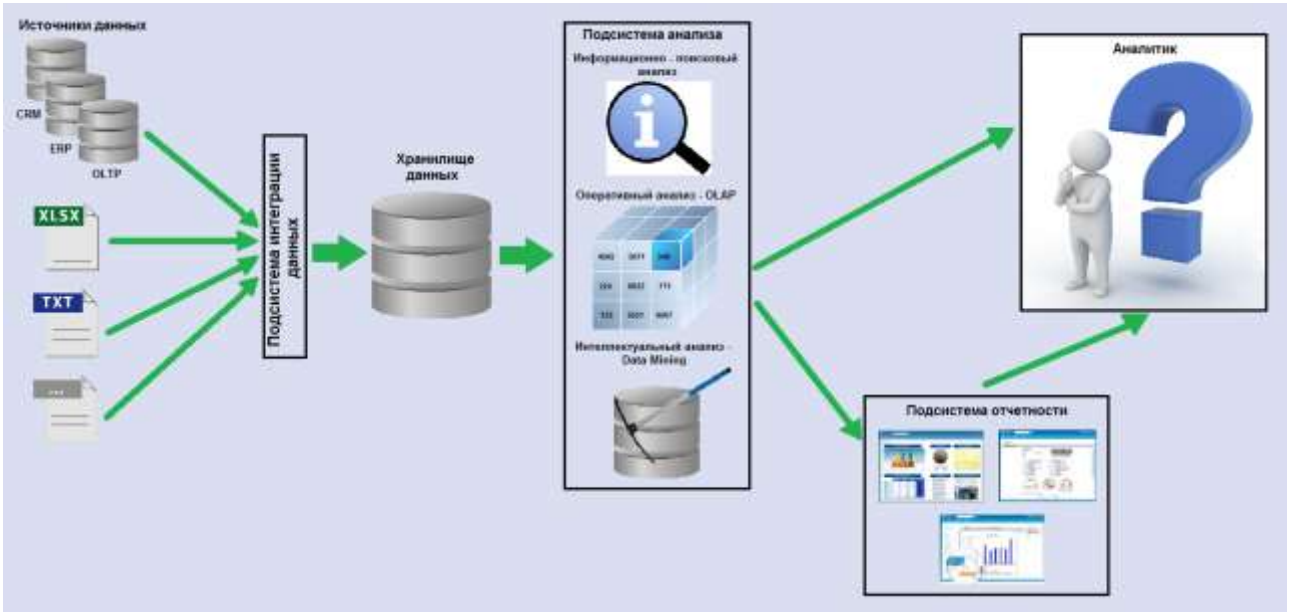

Рис. 1. Обобщенная архитектура системы поддержки принятия решений

Установка программного обеспечения

Во введении приводятся ссылки на интернет страницы, где можно ознакомиться с возможностями, поддерживаемыми различными выпусками SQL Server 2012, минимальными требованиями к оборудованию и операционной системе и ознакомительной версией SQL Server 2012 Enterprise Edition. Здесь описана установка выпуска SQL Server 2012 Enterprise Edition, другие выпуски устанавливаются аналогично.

Откройте **центр установки SQL Server** (Server Installation Center), запустив файл setup.exe с установочного диска. Перед вами появится окно со списком задач в левой части. Первой задачей является **планирование установки SQL Server** (Planning) (рис. 2).

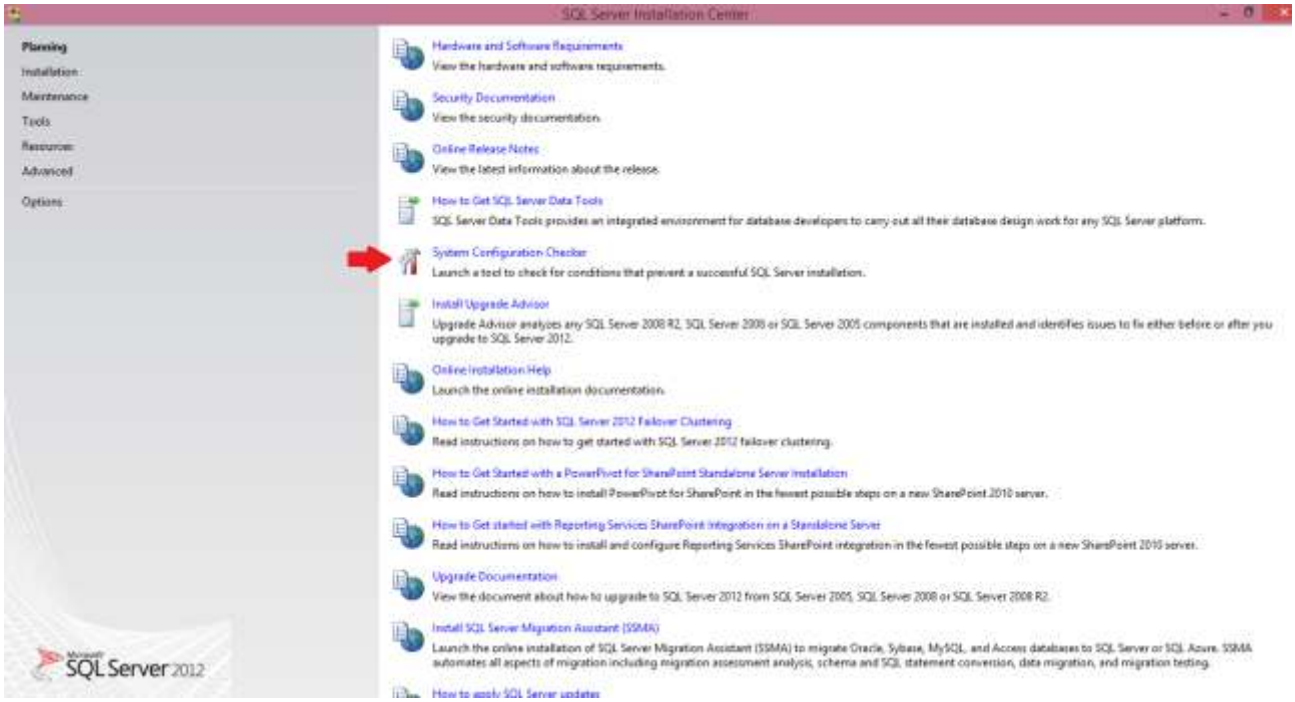

Рис. 2. Центр установки SQL Server Installation Center. Окно планирования

Здесь можно посмотреть требования к оборудованию и программному обеспечению, документацию по безопасности и обновлению продукта и т.д. Важным является пункт **Проверка конфигурации системы** (System Configuration Checker). Выбрав его, вы запустите инструмент для проверки условий, препятствующих успешной установке SQL Server (Setup Support Rules), который определит возможные проблемы при установке вспомогательных файлов SQL Server. По завершению работы выдается отчет (рис. 3). Этот же инструмент запускается в самом начале установки, но все возникшие проблемы следует устранить до её начала.

После проверки можно начать собственно процесс установки, выбрав пункт **Установка** (Installation) из списка слева и ссылку справа **Новая установка изолированного экземпляра SQL Server или добавление экземпляров к существующей установке** (New SQL Server stand-alone installation or add features to and existing installation). Вначале запустится инструмент Setup Support Rules, описанный выше, и если вы запускали его на этапе планирования и устранили все возможные проблемы заранее, то вам не нужно будет прерывать процесс установки. В противном случае необходимо

исправить ошибки и повторить процедуру проверки нажав кнопку **Включить заново** (Re-run).

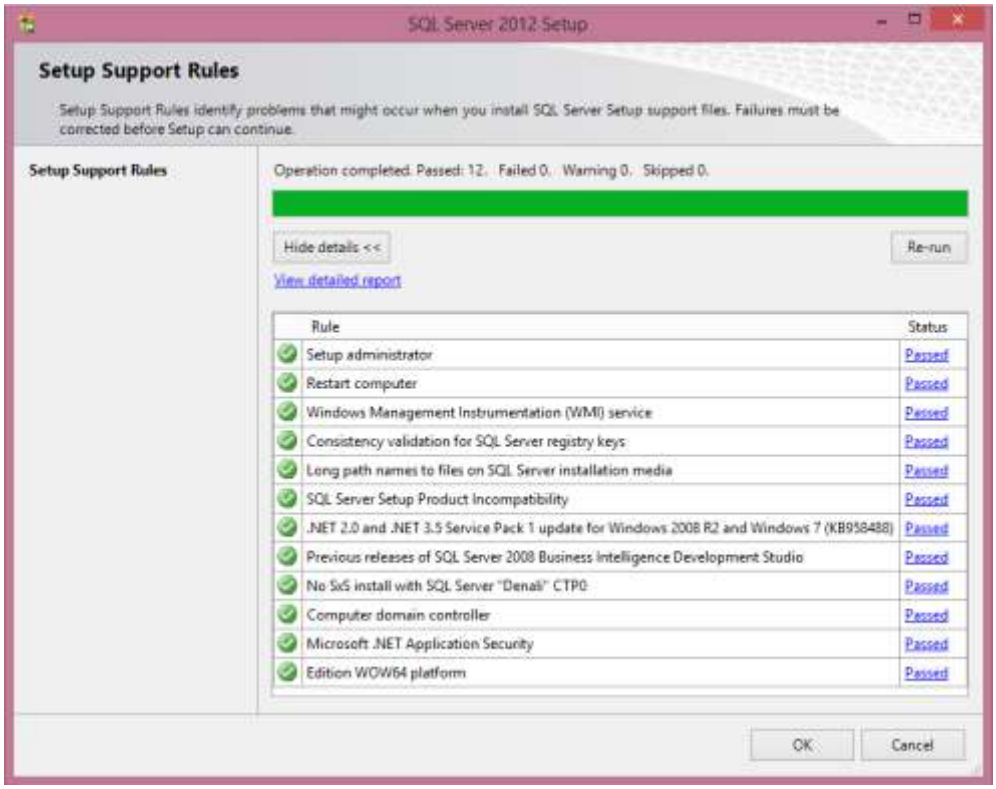

Рис. 3. Отчет о проверке конфигурации

Далее, если у вас имеется 25-значный ключ продукта – его необходимо ввести. Устанавливаемая редакция SQL Server зависит от введенного ключа. Если ключа нет, то можно выбрать пробный выпуск SQL Server – Evaluation Edition с полным набором компонентов и ограничением использования в 180 дней, как показано на рис. 4. Бесплатная версия с ограниченным функционалом – Express Edition для задач бизнес-анализа не подходит.

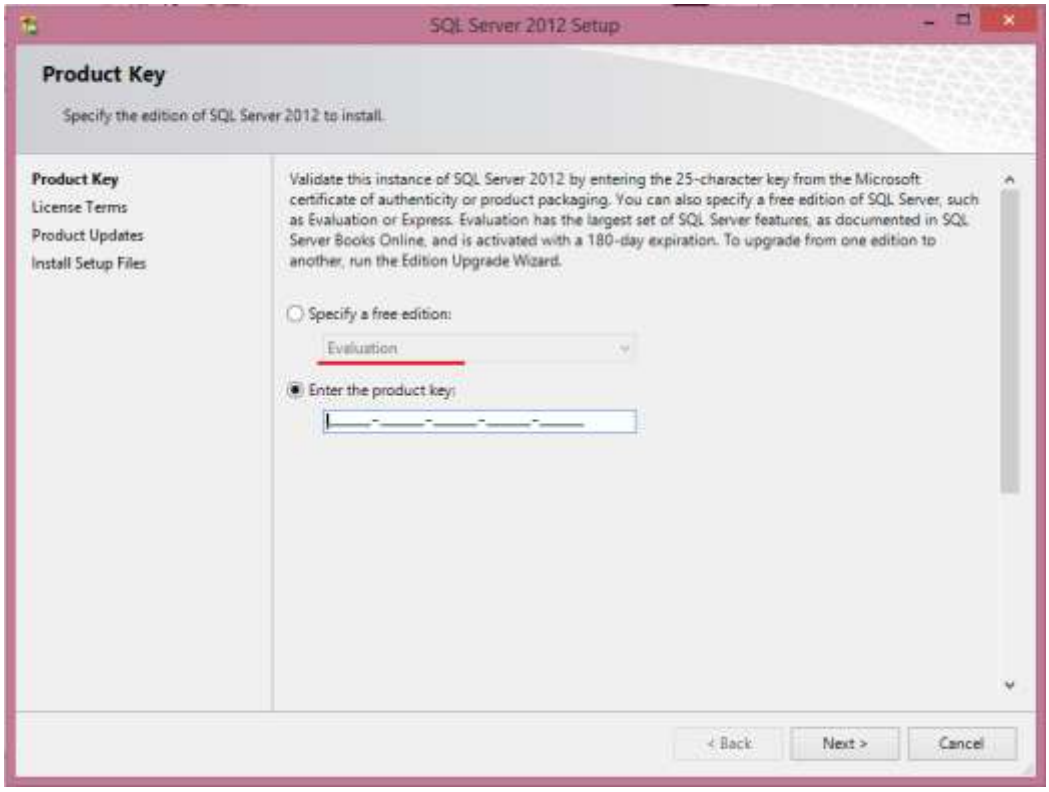

Рис. 4. Выбор редакции SQL Server для установки

Затем необходимо принять условия лицензионного соглашения и решить, нужно ли передавать данные об использовании компонентов в корпорацию Microsoft.

Если имеются обновления программы, то установщик предложит загрузить их из интернета. Рекомендуется всегда включать в установку последние обновления, выбрав **Включить обновления продукта SQL Server** (Include SQL Server product updates) (рис.5).

После этапов загрузки обновлений и инсталляции установочных файлов снова запустится инструмент **Правила поддержки установки** (Setup Support Rules) для определения проблем, которые могут возникнуть во время установки. Все эти проблемы необходимо устранить до начала следующего этапа и повторить процедуру проверки, нажав кнопку **Включить заново** (Re-run) (рис. 6).

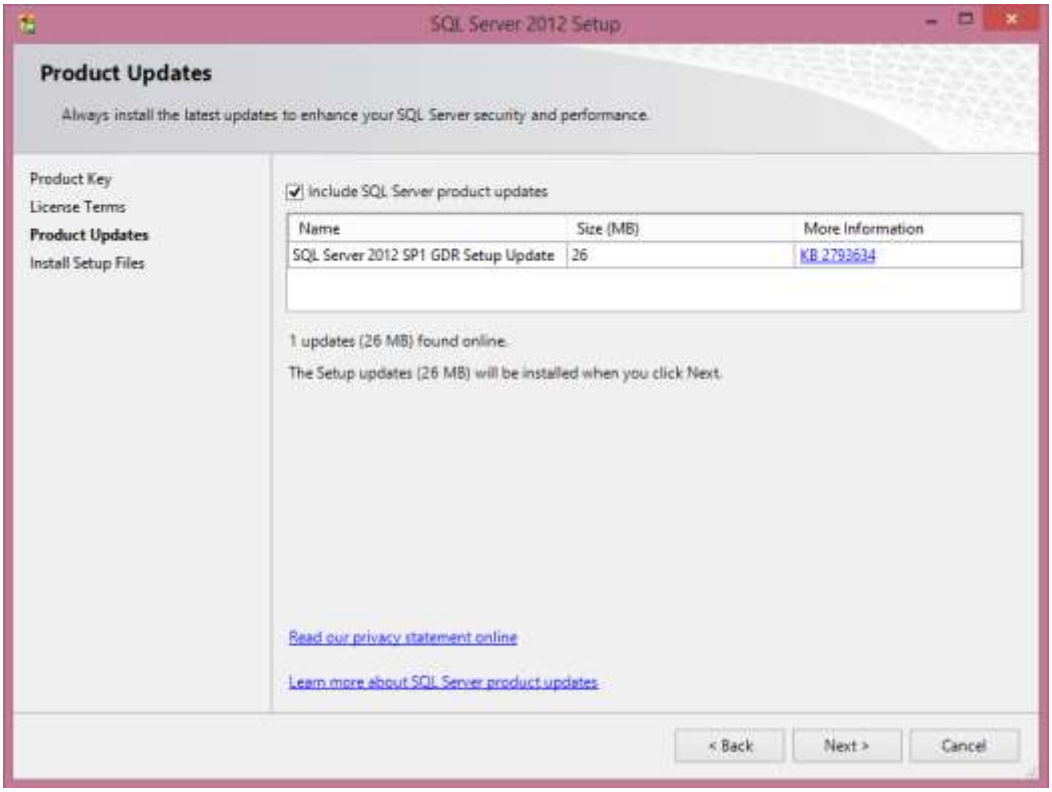

Рис. 5. Поиск и включение последних обновлений перед установкой

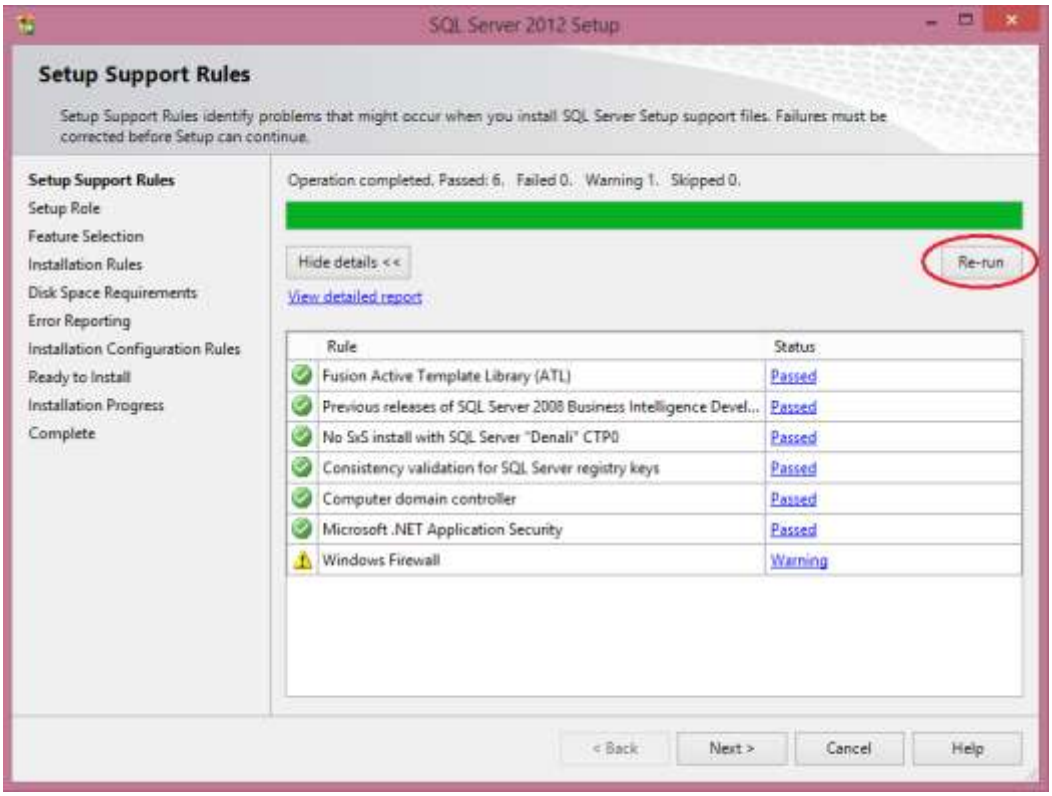

Рис. 6. Поиск возможных проблем во время установки

В следующем окне **Роль установки** (Setup Role) предоставляется возможность выбора установки только основных компонентов SQL Server 2012

(службы компонента Database Engine, службы анализа Analysis Services и службы отчетности Reporting Services), установки дополнительных вспомогательных компонентов PowerPivot для SharePoint или установки всех компонентов со значениями по умолчанию. Выберите первый вариант и нажмите кнопку **Далее** (Next).

Следующим шагом в окне **Выбор компонент** (Feature Selection) необходимо выбрать компоненты, которые требуется установить. На рис. 7 отмечены компоненты, которые потребуются для последующего решения задач бизнес-анализа.

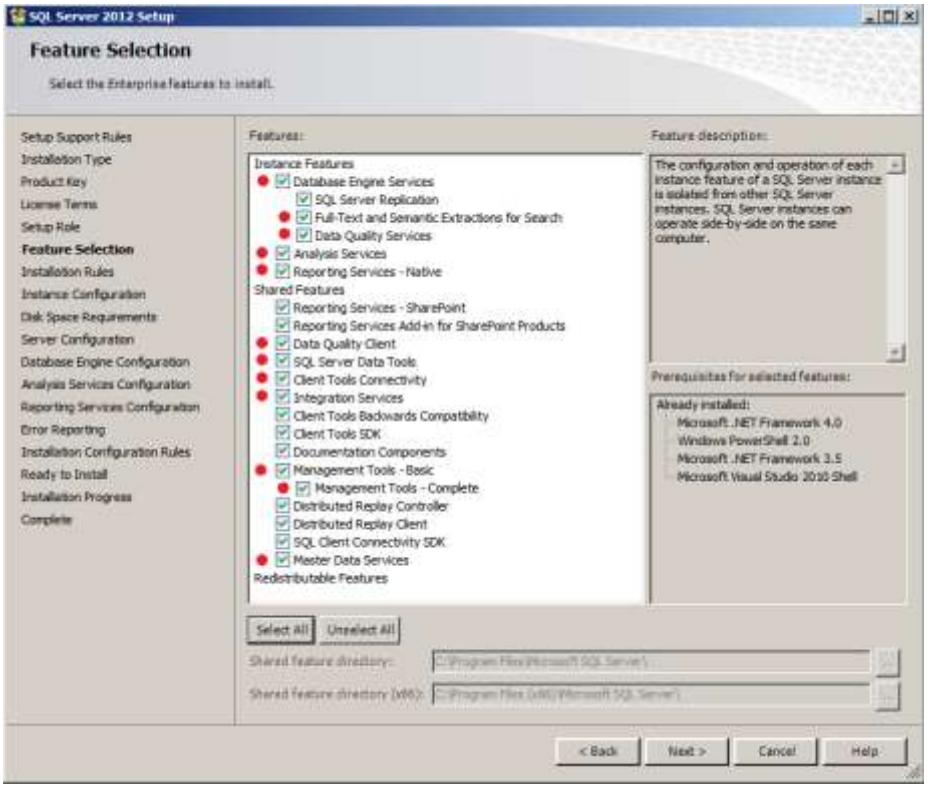

Рис. 7. Выбор компонентов для установки

- ✓ Database Engine Services – основная служба для хранения, обработки и обеспечения безопасности данных, репликации, полнотекстового поиска, средств управления реляционными и XML-данными и сервера служб Data Quality Services (DQS).
- ✓ Data Quality Services – решение, позволяющее поддерживать качество данных и обеспечивать их пригодность к использованию на основе знаний. Эти службы обеспечивают автоматизированные и интерактивные

способы управления целостностью и качеством источников данных. DQS позволяет обнаруживать знания о данных, строить наборы знаний и управлять ими. Затем эти знания можно использовать для выполнения очистки, сопоставления и профилирования данных.

- ✓ Analysis Services – средства создания и управления приложениями оперативной аналитической обработки (OLAP) и приложениями интеллектуального анализа данных.
- ✓ Службы Reporting Services – серверные и клиентские компоненты для создания, управления и развертывания табличных, матричных и графических отчетов, а также отчетов в свободной форме.
- ✓ Службы Integration Services – набор графических средств и программируемых объектов для перемещения, копирования и преобразования данных. Они также включают компонент DQS для служб Integration Services.
- ✓ Среда SQL Server Management Studio – интегрированная среда для доступа, настройки, управления, администрирования и разработки всех компонентов SQL Server.
- ✓ Клиент Data Quality – очень простой и понятный графический пользовательский интерфейс для подключения к серверу DQS и выполнения операций очистки данных.
- ✓ SQL Server Data Tools – среда разработки, интегрированная в Visual Studio и предназначенная для создания решений на основе шаблонов проектов бизнес-аналитики для следующих служб: Analysis Services, Reporting Services и Integration Services.
- ✓ Master Data Services – платформа для интеграции данных из разрозненных систем в масштабе всей организации в единый источник основных данных для целей точности и аудита.

После выбора необходимых компонентов установщик снова запускает проверку системы для успешной установки. В случае отсутствия проблем можно перейти к следующему пункту **Настройка экземпляра** (Instance

Configuration), где необходимо выбрать имя для экземпляра SQL Server или оставить имя экземпляра по умолчанию. На одном изолированном сервере возможно устанавливать до 50 именованных экземпляров MS SQL Server 2012. Экземпляр по умолчанию может быть только один (рис. 8).

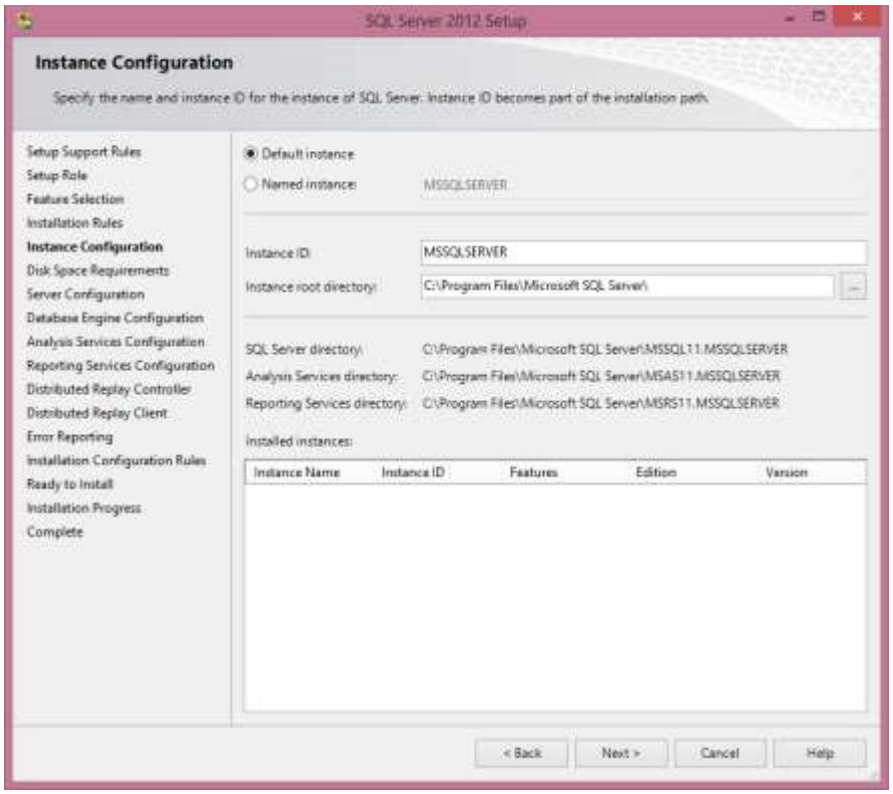

Рис. 8. Именование экземпляра SQL Server

Далее следует проверка требуемого свободного места на диске. После этого следует важный этап **Конфигурирование сервера** (Server Configuration) (рис. 9). Здесь можно задать режимы запуска для устанавливаемых служб SQL Server (возможно позже поменять режим), а также задать учетные записи и пароли, из-под которых они будут запускаться. Можно использовать одну учетную запись для всех служб, но такой подход не рекомендуется по причинам безопасности. В этом же окне на вкладке **Параметры сортировки** (Collation) можно задать порядок сортировки для устанавливаемых компонентов. Для решения задач этого пособия выберите Cyrillic_General_CI_AS для всех компонентов.

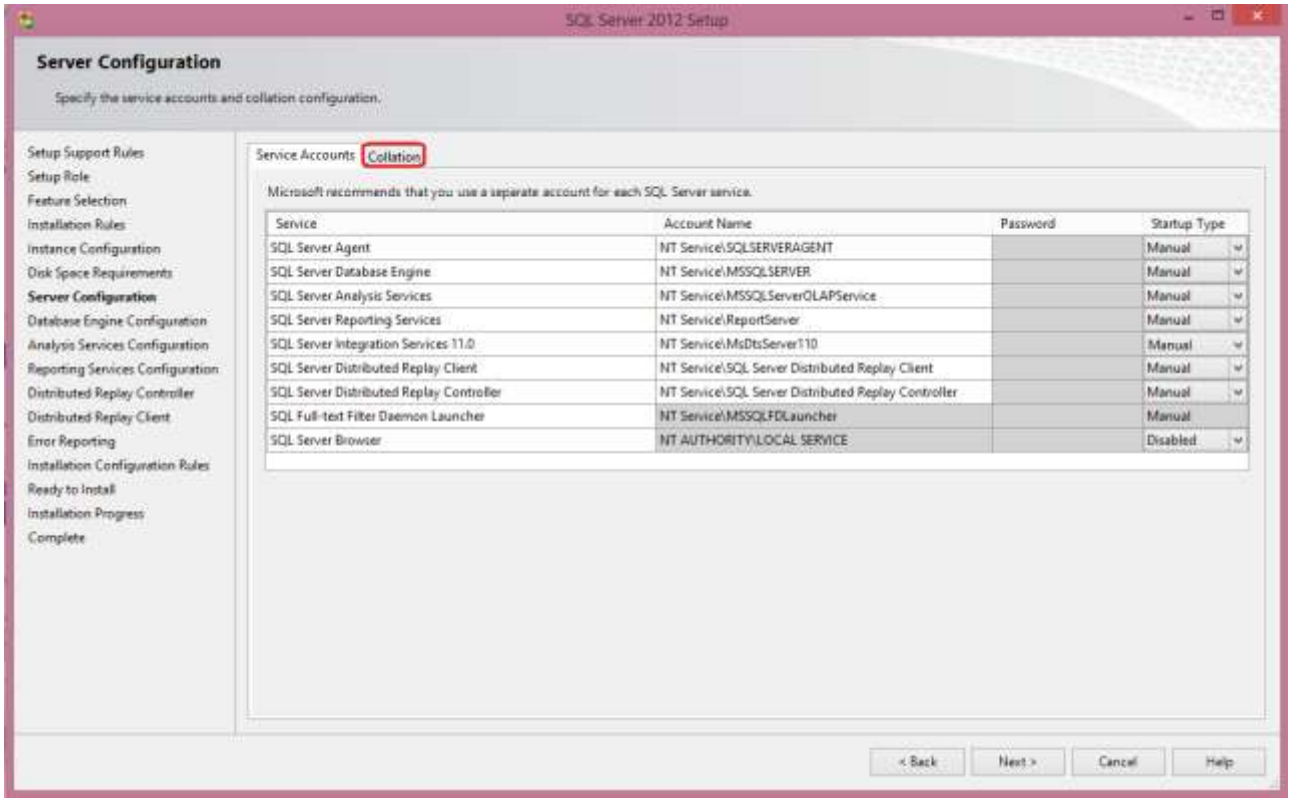

Рис. 9. Конфигурация сервера

Следующим шагом необходимо выбрать режим проверки подлинности для устанавливаемого экземпляра сервера. Компонент Database Engine поддерживает два режима аутентификации: **Режим проверки подлинности Windows** (Windows authentication mode) и **Смешанный режим проверки подлинности SQL Server и Windows** (Mixed Model).

Режим аутентификации Windows является режимом по умолчанию. Вход в SQL Server разрешается определенным группам и пользователям с учетными записями, прошедшими проверку подлинности Windows без предъявления дополнительных учетных данных. Этот режим является более надежным, чем смешанный, так как использует специальный протокол безопасности Kerberos и политику паролей, исходя из проверки сложности надежных паролей, поддерживает блокировку учетных записей и истечение срока пароля.

Поэтому, по возможности, используйте проверку подлинности Windows.

Режим смешанной аутентификации поддерживает проверку подлинности как средствами Windows, так и средствами SQL Server. Пары имен пользователей и паролей отслеживаются в SQL Server [2].

В случае выбора Смешанного режима, необходимо задать и подтвердить пароль для встроенной учетной записи системного администратора SQL Server «sa» (System Administration).

Если был выбран первый вариант, то учетная запись «sa» будет отключена, а пароль будет присвоен программой установки. В случае смены режима на смешанный имя входа «sa» останется отключенным. Его нужно будет включить и задать пароль с помощью команды Transact-SQL – **ALTER LOGIN** или с помощью Management Studio.

Чтобы выбрать режим аутентификации и добавить текущего пользователя в качестве системного администратора к данному экземпляру Database Engine, нажмите кнопку **Добавить текущего пользователя** (Add Current User). Для добавления других пользователей нажмите кнопку **Добавить** (Add) (рис. 10).

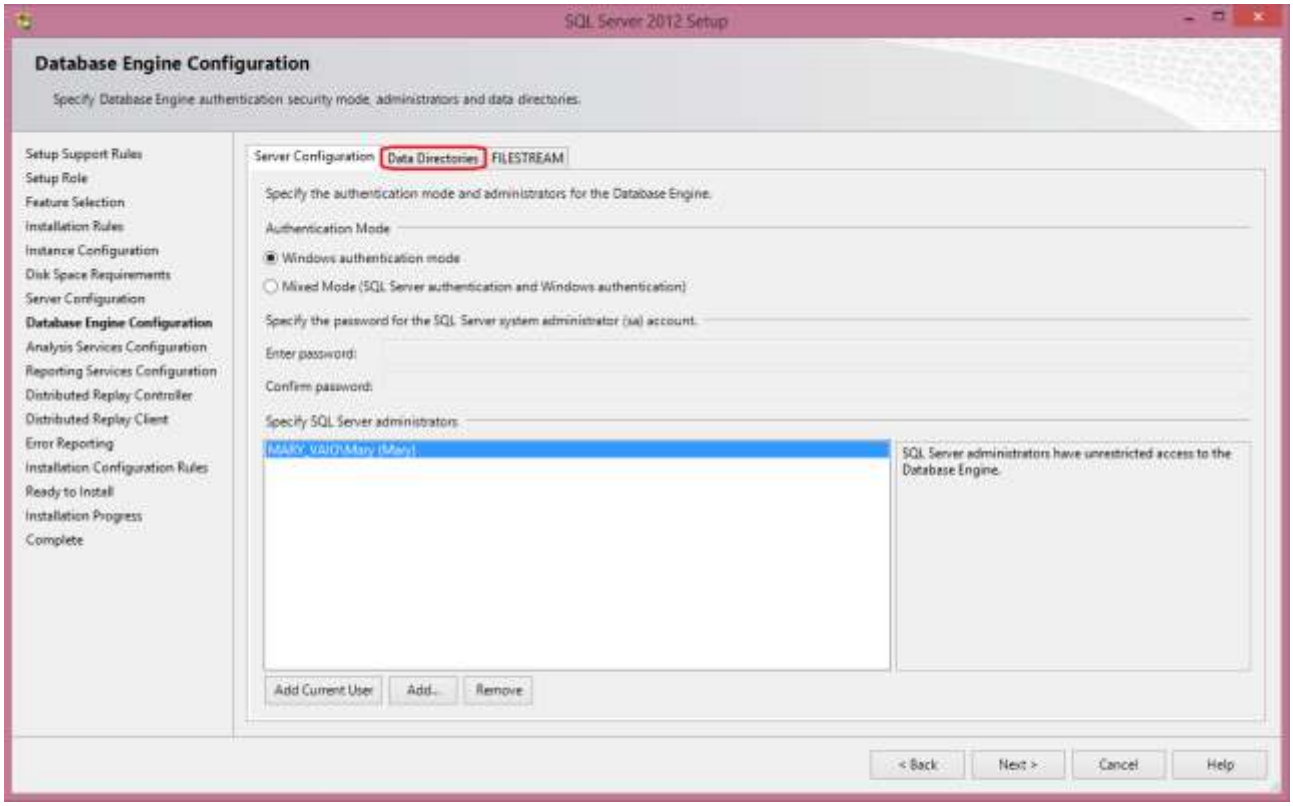

Рис. 10. Выбор режима аутентификации

В этом же окне на вкладке **Каталоги данных** (Data Directories) можно поменять заданное по умолчанию расположение каталогов, в которых будут храниться файлы баз данных, журналов транзакций, резервных копий и др.

Если на данном экземпляре SQL Server планируется хранить

неструктурированные данные, такие как большие документы, изображения, видеофайлы и пр., и для их хранения будет использоваться хранилище FILESTREAM, то его также необходимо настроить на вкладке FILESTREAM.

В следующем окне нужно указать параметры для устанавливаемого экземпляра служб Analysis Services. Здесь можно указать режим сервера, т.е. указать тип баз данных, которые могут быть развернуты на сервере и предоставить права администратора пользователям или службам, которым требуется неограниченный доступ к службам Службы Analysis Services, а на вкладке Data Directories можно поменять заданное по умолчанию расположение каталогов данных (рис. 11).

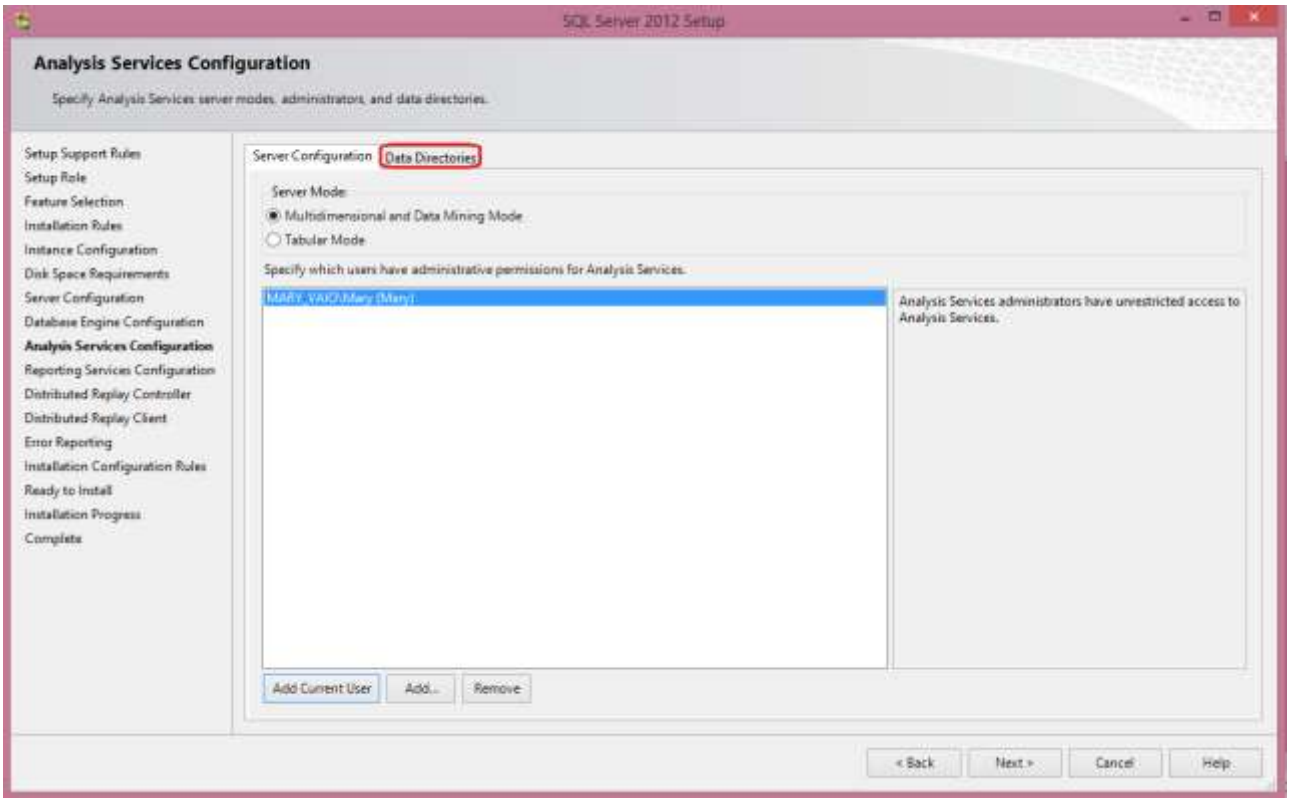

Рис. 11. Выбор параметров установки экземпляра Analysis Services

Службы Analysis Services поддерживают три серверных режима: многомерный и интеллектуальный анализ данных (по умолчанию), PowerPivot для SharePoint (указывается в случае установки служб Analysis Service в составе развертывания PowerPivot для SharePoint) и табличный. Эти режимы различаются архитектурой организации хранилища и использованием памяти. Поэтому на одном экземпляре служб Analysis Services могут запускаться либо

табличные базы данных, либо многомерные базы данных, но не оба типа одновременно. Режим выбирается на этапе установки и не может быть изменен. Если вам потребуется другой режим, необходимо будет дополнительно установить новый экземпляр. В данном курсе рассматривается многомерная модель.

В следующем окне необходимо выполнить **Конфигурирование сервера отчетов** (Reporting Services Configuration), указав параметры установки для этого компонента. Можно установить и настроить его сразу или просто установить, а также интегрировать его с сервером SharePoint.

Выберите просто установку (рис. 12). Подробное описание настройки сервера отчетов будет дано позже.

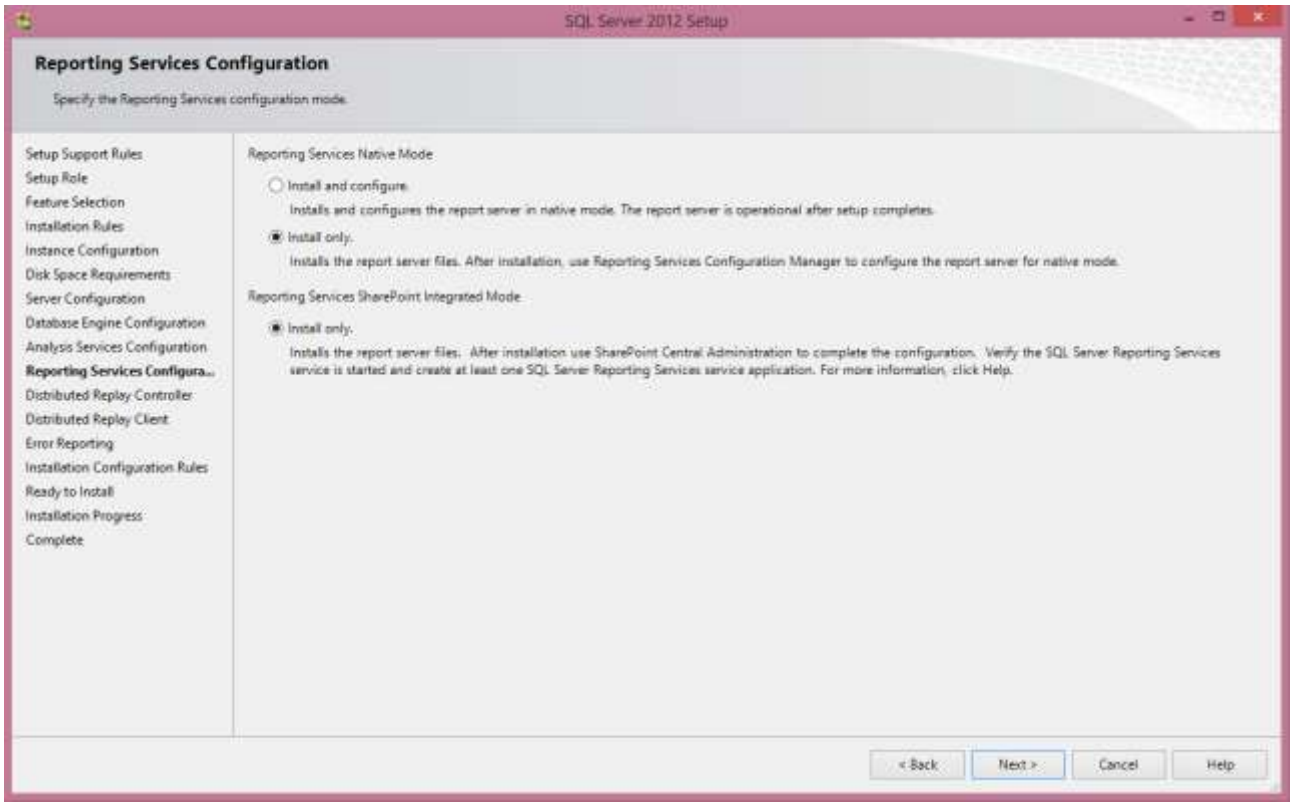

Рис. 12. Выбор параметров установки служб Reporting Services

Следующим шагом в окне **Отчет об ошибках** (Error Reporting) можно указать автоматическую отправку отчета об ошибках сервера компании Microsoft.

Далее следует последняя проверка конфигурации перед установкой и

отчет о ней выдается в окне **Правила конфигурации установки** (Installation Configuration Rules).

Окно **Все готово для установки** (Ready to Install) является последним перед её началом. Здесь можно просмотреть сводную информацию об устанавливаемых компонентах SQL Server. Чтобы начать процесс установки нажмите кнопку **Установить** (Install) в этом окне.

Глава 1. Концепция хранилищ данных

Средства бизнес-анализа не обязательно должны включать в себя компонент хранилища данных, но в этом случае проблемы сбора, очистки и согласования данных должны решаться либо в режиме реального времени, либо заранее для определенного круга запросов, что не может не сказаться на производительности как системы бизнес-анализа, так и на базах оперативной обработки транзакций (OLTP-системы, от анг. Online Transaction Processing), откуда эти данные берутся. Это связано с противоречивостью требований, предъявляемых к этим системам, а также со значительными сложностями в обеспечении эффективности выполнения аналитических запросов и обслуживании оперативной деятельности пользователей в многопользовательской среде.

Рассмотрим основные различия в требованиях, предъявляемых к таким системам и проблемы разграничения оперативной и аналитической деятельности.

Частота изменений данных и временные интервалы хранения

OLTP-системы, как правило, имеют дело с текущими значениями каких-либо показателей. Данные в них хранятся за относительно небольшой период времени и постоянно изменяются. Поэтому основное требование к таким системам – эффективное выполнение операций по модификации базы данных (добавление, изменение, удаление данных) и быстрый поиск отдельных строк данных.

Для анализа данных, наоборот, необходимы сведения за максимально большой период времени. В отличие от OLTP-систем, BI-системы хранят исторические данные за годы работы предприятия и единожды попав в систему, уже практически не изменяются. Ввод новых данных носит, как правило, эпизодический характер и выполняется в период низкой активности системы, а основной операцией в хранилище является выборка большого числа строк данных.

Избыточность данных

Структура базы данных OLTP-системы обычно высоко нормализована. Она может содержать десятки и даже сотни таблиц, ссылающихся друг на друга. Это позволяет в значительной степени исключить избыточность данных и, как следствие, противоречивость при выполнении коротких обновляющих транзакций (операций вставки, удаления или изменения записей в таблицах), а также оптимизировать выполнение выборки данных, затрагивающей только отдельные записи при работе в многопользовательской среде.

Аналитические запросы, как правило, выполняются над большим количеством данных с широким применением группировок и обобщений. В этом случае высокая нормализация замедляет выполнение запроса, так как требуется выполнять множество операций по объединению таблиц, поэтому при проектировании систем анализа стараются максимально упростить схему базы данных и уменьшить количество таблиц, участвующих в запросе. С этой целью часто допускают *денормализацию* (избыточность данных).

Качество данных

OLTP-системы, как правило, хранят информацию, вводимую непосредственно пользователями системы, что повышает вероятность ввода ошибочных данных и может создавать локальные проблемы в системе.

В аналитические системы должны попадать уже выверенные данные во избежание неправильных выводов и принятия неверных стратегических решений.

Формат хранения данных

OLTP-системы, обслуживающие различные участки работы, могут быть не связаны между собой. Они часто реализуются на разных программно-аппаратных платформах. Одни и те же данные в разных базах могут быть представлены в различном виде и не совпадать. В процессе анализа такое различие форматов чрезвычайно затрудняет совместный анализ данных. Поэтому к системам анализа предъявляется требование единого формата и, как правило, необходимо чтобы этот формат был оптимизирован для анализа

данных.

Проблемы разграничения оперативной и аналитической деятельности в многопользовательской среде

При выполнении аналитических запросов в многопользовательской среде накладываются дополнительные блокировки, что снижает производительность как оперативных, так и аналитических запросов за счет дополнительной конкуренции.

Как правило, аналитические запросы довольно сложны и выполняются дольше, чем операционные. Это может привести к несогласованности данных в аналитическом отчете, так как за время его построения данные могут измениться в ходе операционной деятельности.

Для небольших компаний с невысокой интенсивностью аналитической деятельности можно строить компромиссные решения путем использования различных механизмов разграничения деятельности, однако, если в компании интенсивность аналитической деятельности высока и есть необходимость строить отчеты в режиме реального времени, то построить высокопроизводительное решение на базе OLTP-системы не удастся.

В настоящее время наиболее популярным решением этой проблемы является подход, ориентированный на использование концепции *хранилищ данных*.

Общая идея заключается в разделении баз данных для OLTP-систем и систем анализа и последующем их проектировании с учетом соответствующих требований. Такое разделение позволяет разработать оптимизированную структуру данных транзакционной базы и структуру, используемую для анализа.

Автором концепции хранилищ данных является Б. Инмон, который определил их как предметно-ориентированные, интегрированные, неизменчивые, поддерживающие хронологию наборы данных, организованные для целей поддержки принятия решений. Они призваны выступать в качестве централизованного склада согласованных бизнес-данных для последующего

анализа и создания отчетов, позволяющие разгрузить оперативные базы данных. Это реляционные базы данных, но спроектированные с учетом повышения производительности аналитических запросов.

Рассмотрим свойства хранилищ данных более подробно.

Предметная ориентированность. Все данные, отражающие разные точки зрения на одну предметную область, собираются (обычно из множества различных оперативных источников данных), очищаются, согласовываются, дополняются, агрегируются и представляются в единой, удобной для их анализа форме. Предметная ориентированность позволяет также хранить в ХД только те данные, которые нужны для анализа.

Интегрированность. Все данные взаимно согласованы (приведены к единому формату) и хранятся в едином общекорпоративном хранилище.

Поддержка хронологии. Для анализа данных очень важно иметь возможность отслеживать хронологию изменений показателей предметной области, поэтому данные хронологически структурированы и отражают историю за достаточный период времени, для выполнения задач анализа и прогнозирования.

Неизменяемость. Исходные (исторические) данные, после того как они были согласованы, верифицированы и внесены в общекорпоративное хранилище, остаются неизменными и используются исключительно в режиме чтения.

Существует и ряд проблем, связанных с использованием хранилищ данных при построении аналитических систем. Самая очевидная проблема – это усложнение всей структуры BI-системы и, как следствие, её удорожание. Нужно обслуживать различные базы, поддерживать процесс переноса данных и т.д. Другая проблема заключается в том, что данные в ХД не совсем актуальны, так как попадают туда из оперативных источников с некоторым опозданием.

Тем не менее, преимущества от использования хранилищ при построении систем бизнес-анализа очевидны.

Так как хранилища данных оперируют огромными объемами

информации, то к их проектированию и реализации предъявляют повышенные требования. Хотя конкретные детали отдельных решений могут варьироваться, есть некоторые общие приемы в большинстве реализаций хранилищ данных. Знакомство с этими приемами позволит лучше спланировать и построить эффективное решение.

При создании ХД следует учитывать следующие требования.

- ✓ Необходимость интеграции данных из неоднородных источников в распределенной среде.
- ✓ Потребность в эффективном хранении и оптимизации операций чтения при обработке очень больших объемов информации.
- ✓ Загрузка новых или обновленных данных через регулярные промежутки времени.
- ✓ Необходимость многоуровневых справочников метаданных. Для систем анализа наличие развитых метаданных (данных о данных) и средств их представления конечным пользователям является одним из основных условий успешной реализации ХД. Метаданные необходимы пользователям для понимания структуры информации, на основании которой принимается решение.
- ✓ Повышенные требования к безопасности данных. В ХД хранится конфиденциальная информация, доступ к которой ограничен в пределах организации, не говоря уже о других организациях.

Существует несколько типичных подходов к архитектуре хранилищ данных.

- ✓ Создание единого центрального хранилища предприятия для всех бизнес-единиц.
- ✓ Создание небольших ведомственных витрин данных (Data Mart), то есть тематических БД, содержащих информацию, относящуюся к отдельным аспектам деятельности организации.
- ✓ Создание хранилища по схеме «звезда» с синхронизацией центрального

склада данных предприятия с ведомственными витринами данных, которые содержат подмножества данных из центрального хранилища данных.

На рис. 13 показаны типичные варианты архитектуры ХД.

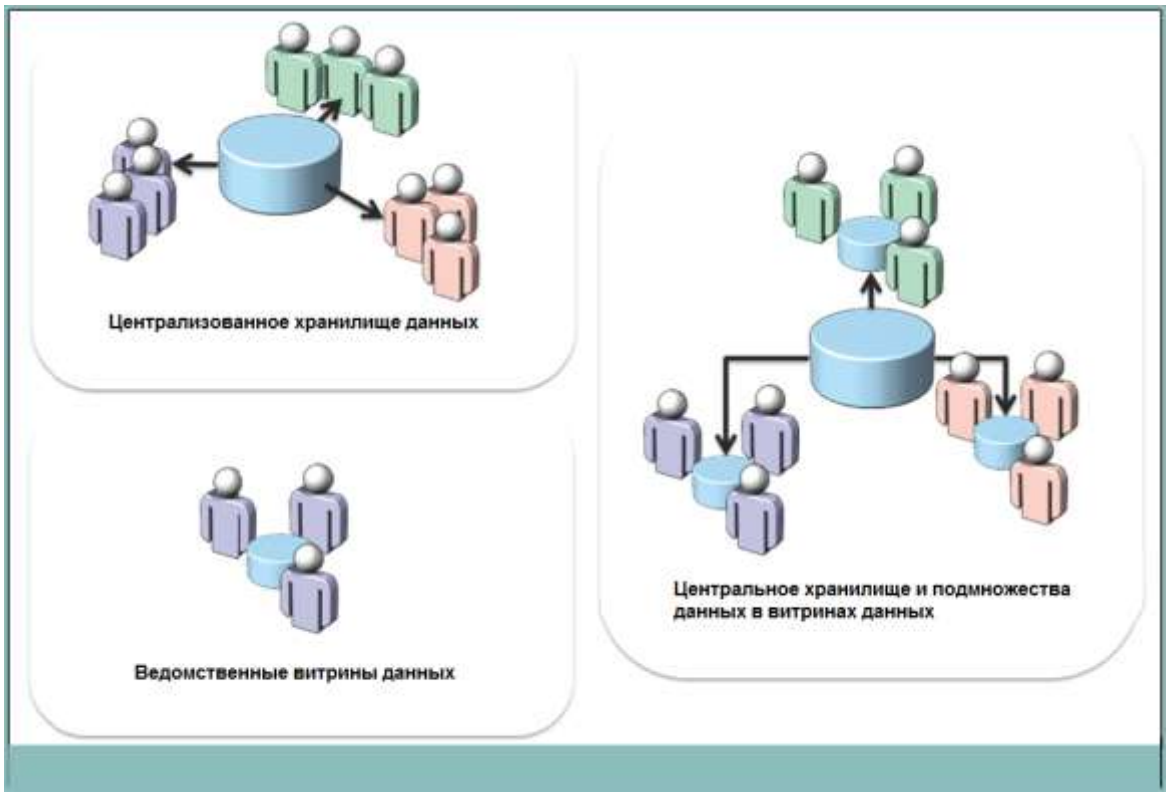

Рис. 13. Архитектура хранилища данных

На рис. 14 представлена инфраструктура хранилища данных, которая обычно состоит из следующих компонентов:

Источники данных. Как правило, на предприятиях существуют несколько баз данных, обслуживающих различные участки работы. Часть данных может находиться в электронных таблицах, текстовых файлах, загружаться с Web-серверов, и т.д. Поэтому первоочередной задачей является консолидация данных из различных источников.

ETL-процесс (от анг. Extract, Transform, Load). Процесс доступа к различным источникам данных, их извлечения, преобразования и загрузки в хранилище. На этом этапе данные очищаются, приводятся к единому формату, производятся необходимые вычисления для того, чтобы они соответствовали модели данных хранилища.

Временное хранилище. Часто данные не сразу загружаются в основное хранилище, а сначала попадают в промежуточную базу – посредник. Например, она может потребоваться в тех случаях, когда необходимо синхронизировать данные, загружаемые из различных источников (доступ к источникам происходит в разное время) или требуется провести сложную предварительную обработку данных.

Хранилище данных. Реляционная база данных, спроектированная для обеспечения высокой производительности запросов к историческим бизнес-данным для анализа и создания отчетности.

Часто инфраструктура ХД включает в себя дополнительные компоненты:

Очистка данных и удаление дубликатов для повышения качества данных, прежде чем они будут загружены в хранилище.

Управление нормативно-справочной информацией. Это совокупность процессов и инструментов для приведения данных из разных систем, используемых на предприятии, к единому стандарту. Для этого используется мастер-справочник и вся поступающая информация сопоставляется с данными в справочнике.

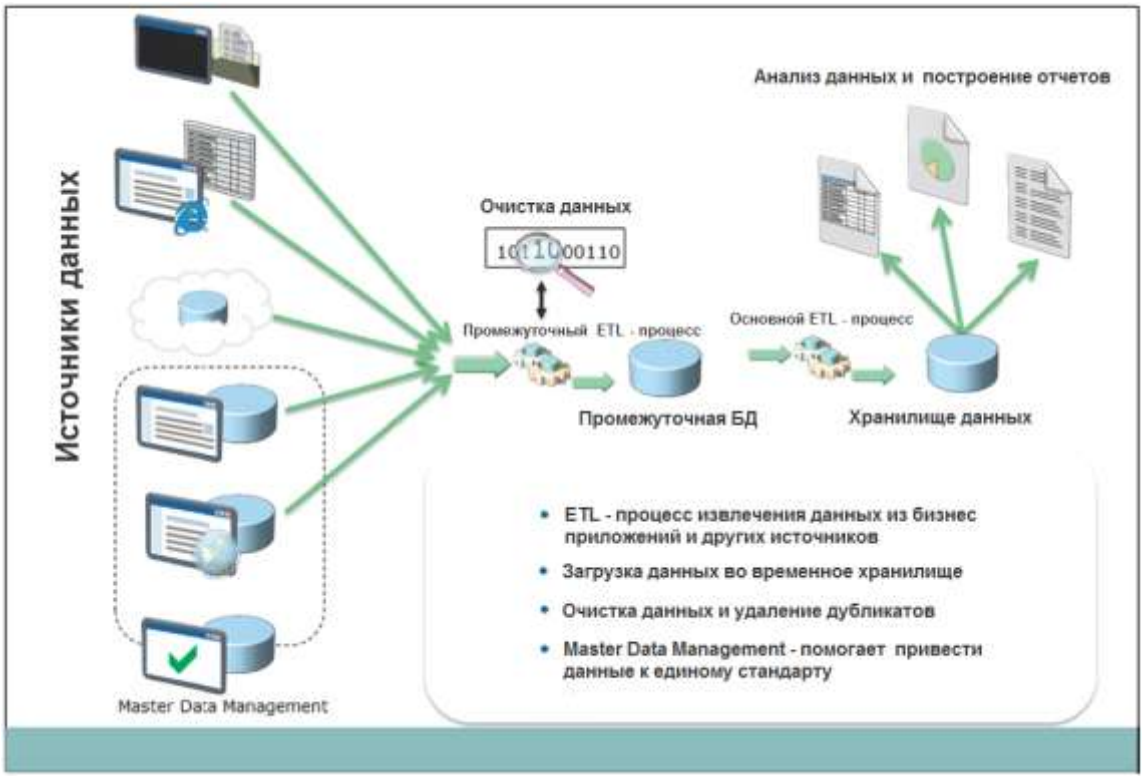

Рис. 14. Типичная инфраструктура хранилища данных

Контрольные вопросы

1. Какие проблемы возникают при анализе данных непосредственно в транзакционной базе?
2. Почему в хранилище данных допускается денормализация?
3. Перечислите основные свойства хранилищ данных.
4. Какие существуют варианты архитектуры хранилищ данных?
5. Каким образом данные попадают в хранилище?
6. Что такое ETL-процесс?
7. Для чего нужно промежуточное хранилище данных?
8. Для чего нужна очистка данных?
9. Что такое Master Data Management?