

Лабораторная работа №2

Современные технологии анализа данных

Цель работы: изучить современные инструменты для анализа данных на примере Microsoft Analysis Services, Deductor Studio 5.2.

Задачи работы:

- изучить краткую теорию;
- выполнить задание и защитить отчет у преподавателя.

1. Краткая теория

Data Mining (интеллектуальный анализ данных) – это процесс извлечения действительной, подлинной информации из больших баз данных. Другими словами, интеллектуальный анализ данных находит некоторые тенденции, которые существуют в данных. Эти закономерности и тенденции называются моделью интеллектуального анализа данных, которые могут быть применены для прогнозирования продаж, определения комплементарных товаров, поиск последовательностей в порядке выбора клиентами товаров.

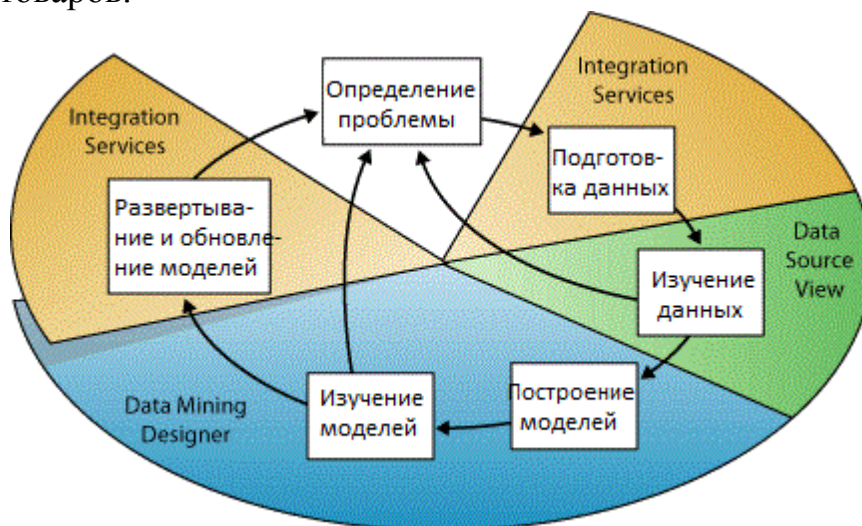


Рисунок 1 – Процесс построения модели интеллектуального анализа данных

Построение модели интеллектуального анализа данных делится на шесть этапов: определение проблемы, подготовка данных, изучение данных, построение моделей, изучение и проверка моделей, развертывание и обновление моделей.

Термин «Business Intelligence» был введен в обращение аналитиками компании Gartner Group в конце 80-х годов прошлого века как пользовательцентрический процесс, включающий доступ и исследование информации, ее анализ, выработку интуиции и понимания, которые ведут к улучшенному и неформальному принятию решений. Хотя ранее этот термин, например, использовался в компании IBM в качестве внутрикорпоративного термина.

Business Intelligence так же понимают как инструменты для анализа данных, построения отчетов и запросов, которые могут помочь бизнес-пользователям преодолеть море данных для того, чтобы помочь синтезировать из них значимую информацию.

Системы Business Intelligence – это совокупность технологий, программного обеспечения и практик, направленных на достижение целей бизнеса путем наилучшего использования имеющихся данных.

Среда Business Intelligence Development Studio является версией Microsoft Visual Studio, адаптированную для работы с бизнес-аналитикой в SQL Server (рисунок 2), она используется для создания новых OLAP-кубов.

Для того, чтобы ознакомиться со средой Business Intelligence Development Studio выполните: Пуск > Все программы > Microsoft SQL Server 2008 > Среда SQL Server Business Intelligence Development Studio.

ВНИМАНИЕ: В зависимости от версии MS SQL Server путь может отличаться, лабораторные курсы выполнены с использованием Microsoft SQL Server 2008.

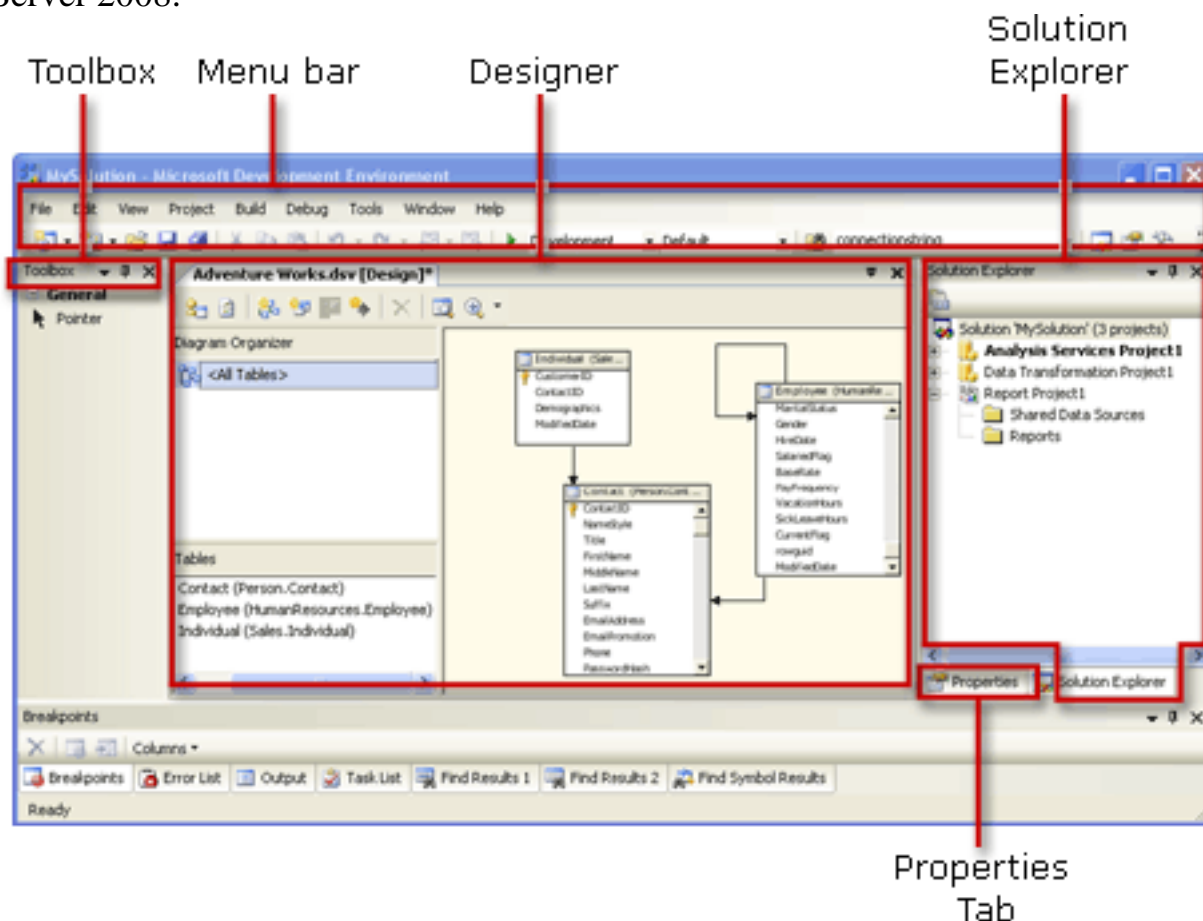


Рисунок 2 – Интерфейс среды Business Intelligence Development Studio

Microsoft Analysis Services – это средство для анализа данных посредством Business Intelligence и Data Mining, являющееся частью Microsoft SQL Server. В службах Службы Analysis Services предусмотрен ряд решений для построения и развертывания аналитических баз данных, используемых для поддержки принятия решений в Excel, PerformancePoint, службах Reporting Services и других приложениях бизнес-аналитики. Базой

любого решения Analysis Services являются семантическая модель данных бизнес-аналитики и экземпляр сервера, на котором создаются, обрабатываются, отправляются запросы и производится управление объектами в этой модели. В составе Analysis Manager имеется простейшее средство просмотра многомерных данных, представляющее собой элемент управления ActiveX, использующий для доступа к данным OLE DB for OLAP.

Analysis Manager использует библиотеки SQL DSO для создания и модификации объектов многомерной базы данных и OLE DB для доступа к исходным реляционным хранилищам данных. Что касается доступа к самим многомерным данным, то Analysis Manager использует для этой цели OLE DB for OLAP.

Microsoft Analysis Services устанавливаются при установке Microsoft SQL Server, либо могут быть установлены в качестве отдельного компонента.

Логическая архитектура.

Экземпляр служб Analysis Services может содержать несколько баз данных, а в базе данных могут одновременно присутствовать объекты OLAP и объекты интеллектуального анализа данных. Приложения подключаются к указанному экземпляру служб Analysis Services и к указанной базе данных. На серверном компьютере может эксплуатироваться несколько экземпляров служб Analysis Services. Экземпляры служб Analysis Services именуются как «<ИмяСервера>\<ИмяЭкземпляра>». В следующей иллюстрации показаны все упомянутые связи между объектами служб Analysis Services.

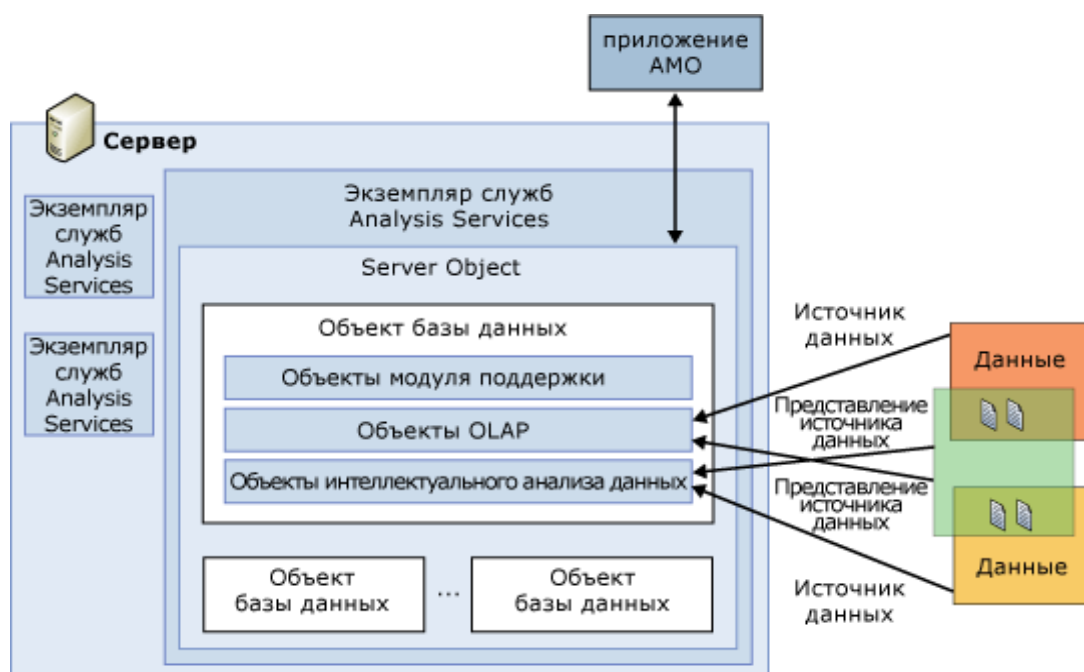


Рисунок 3 – Связи между объектами

Основные классы представляют собой минимальный набор объектов, требуемый для формирования куба. Этот минимальный набор объектов включает измерение, группу мер и секцию. Определение статистической обработки является необязательным.

Измерения создаются на основе атрибутов и иерархий. Иерархии формируются с использованием упорядоченного набора атрибутов, такого, что каждый атрибут соответствует одному из уровней в иерархии.

Кубы создаются на основе измерений и групп мер. Измерения в коллекции измерений куба принадлежат к коллекции измерений базы данных. Группы мер — это коллекции мер, которые имеют одно и то же представление источника данных и одно и то же подмножество измерений в кубе. Группа мер имеет одну или несколько секций, предназначенных для управления физическими данными. Группа мер может иметь применяемую по умолчанию статистическую схему. Статистическая схема по умолчанию может использоваться во всех секциях в группе мер; кроме того, каждая секция может иметь собственную статистическую схему.

Объекты сервера

Каждый экземпляр служб Analysis Services рассматривается как отдельный объект сервера среди объектов АМО; каждый отдельный экземпляр подключается к объекту Server с помощью отдельного соединения. Каждый объект сервера содержит один или несколько источников данных, представление источника данных и объекты базы данных, а также сборки и роли безопасности.

Объекты измерений

Каждый объект базы данных содержит несколько объектов измерения. Каждый объект измерения содержит один или несколько атрибутов, которые организованы в виде иерархий.

Объекты куба

Каждый объект базы данных содержит один или несколько объектов куба. Куб задается его мерами и измерениями. Меры и измерения куба выводятся из таблиц и представлений в представлении источника данных, на котором основан куб или который создан из определений мер и измерений.

Объектная модель ASSL содержит много повторяющихся групп элементов. Например, группа элементов «Dimensions contain Hierarchies» определяет иерархию измерений элемента. И в кубах Cubes, и в группах мер MeasureGroups содержится группа элементов «Dimensions contain Hierarchies».

Если элемент не переопределяется явно, он наследует все характеристики этой повторяющейся группы элементов от более высокого уровня. Например, значения Translations для измерения куба CubeDimension являются такими же, как и значения Translations для элемента Cube, являющегося его предком.

Чтобы иметь возможность явно переопределять свойства, унаследованные от объекта более высокого уровня, не обязательно явно повторять в объекте всю структуру и свойства объекта высокого уровня. Единственными свойствами, которые требуют явного определения в объекте, являются те свойства, которые необходимо переопределить. Например, в измерении куба CubeDimension могут быть перечислены только те иерархии Hierarchies, которые должны быть запрещены для использования в кубе

Cube, или те, что требуют изменения признака видимости, или же те, для которых некоторые подробности Level не были предусмотрены на уровне Dimension.

Некоторые свойства, заданные в объекте, предоставляют применяемые по умолчанию значения для дочерних объектов или объектов-потомков. Например, свойство Cube.StorageMode предоставляет значение, применяемое по умолчанию для свойства Partition.StorageMode. Что касается унаследованных значений по умолчанию, то в языке ASSL применяются такие же правила, что и для объектов DSO 8.0. В следующем списке приведены правила, касающиеся унаследованных значений по умолчанию.

Если свойство для дочернего объекта не определено в элементе XML, то по умолчанию в качестве значения свойства применяется унаследованное значение. Но, если выполняется запрос этого значения с сервера, сервер возвращает неопределенное значение элемента XML.

Возможность определить программным путем, было ли задано свойство дочернего объекта непосредственно на дочернем объекте или унаследовано, отсутствует.

Пример.

Куб «Импорт» содержит две меры («Пакеты» и «Последняя дата») и три связанных измерения («Маршрут», «Источник» и «Время»).

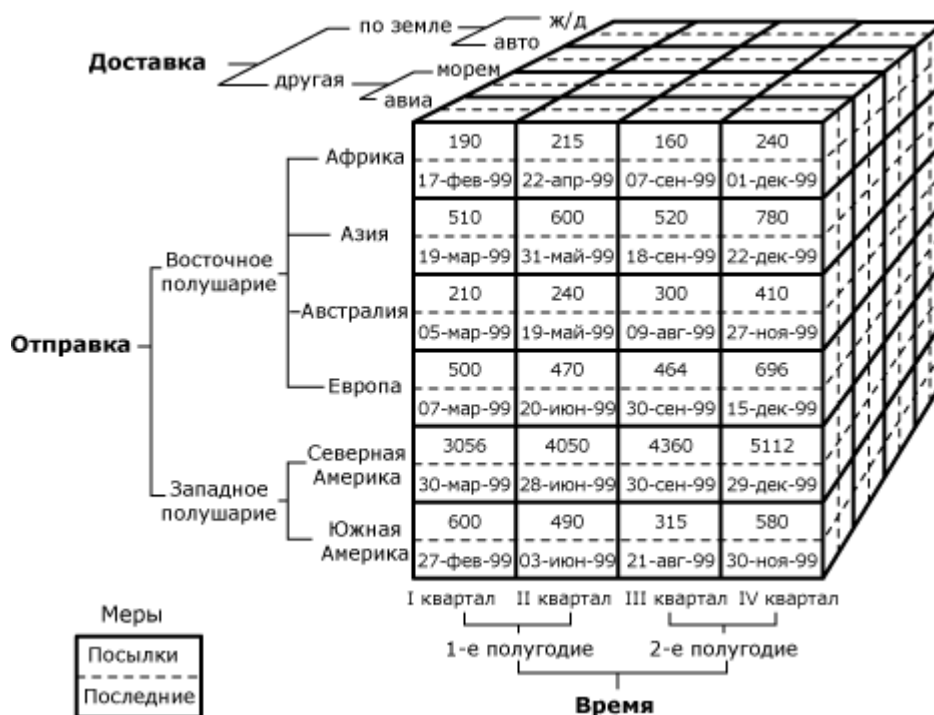


Рисунок 4 – Куб данных

Наименьшие буквенно-цифровые значения в кубе — это элементы измерений. Примеры элементов — «Наземный» (элемент измерения «Маршрут»), «Африка» (элемент измерения «Источник») и «1-й квартал» (элемент измерения «Время»).

Меры

Значение в ячейках куба представляют две меры — «Пакеты» и «Последняя дата». Мера «Пакеты» представляет число импортированных посылок; для статистической обработки фактов используется функция Sum. Мера «Последняя дата» представляет собой дату получения; для статистической обработки используется функция Max.

Измерения

Измерение «Маршрут» представляет пути, которыми импортируемый товар достигает своего назначения. В число элементов этого измерения входят «наземный», «не наземный», «воздушный», «морской», «дорожный» и «железнодорожный». Измерение «Источник» представляет место производства импортируемого товара, например Азию или Африку. Измерение «Время» представляет кварталы и полугодия.

Статистические вычисления

Бизнес-пользователи куба могут определять значения его мер для каждого элемента в каждом измерении независимо от уровня элемента в измерении, поскольку службы Analysis Services вычисляют значения верхних уровней по мере необходимости. Например, значения меры в предыдущей иллюстрации могут быть вычислены в соответствии с обычной календарной иерархией с использованием иерархии «Календарное время» в измерении «Время», как показано на следующей диаграмме.



Рисунок 5 – Пример использования иерархии «Календарное время» в измерении «Время»

Помимо статистических вычислений с использованием одного измерения, возможна статистическая обработка мер с использованием сочетаний элементов из различных измерений. Это позволяет бизнес-пользователям вычислять меры в нескольких измерениях одновременно. Например, если бизнес-пользователь хочет проанализировать квартальный

импорт, прибывший по воздуху из восточного и западного полушарий, он может создать запрос к кубу, чтобы получить следующий набор данных.

			Посылки		Последняя дата			
			Все источники	Восточное полушарие	Западное полушарие	Все источники	Восточное полушарие	Западное полушарие
Все время			25110	6547	18563	29-дек-99	29-дек-99	29-дек-99
	Первое полугодие		11173	2977	8196	28-июн-99	28-июн-99	28-июн-99
		Первый квартал	5108	1452	3656	30-мар-99	30-мар-99	30-мар-99
		Второй квартал	6065	1525	4540	28-июн-99	28-июн-99	28-июн-99
	Второе полугодие		13937	3570	10367	29-дек-99	29-дек-99	29-дек-99
		Третий квартал	6119	1444	4675	30-сен-99	30-сен-99	30-сен-99
		Четвертый квартал	7818	2126	5692	29-дек-99	29-дек-99	29-дек-99

После определения куба можно создать новые агрегаты или изменить существующие агрегаты, установив параметры наподобие того, вычисляются ли агрегаты предварительно во время обработки или же вычисляются во время запроса.

Сопоставление мер, атрибутов и иерархий

Меры, атрибуты и иерархии в примере куба выводятся из следующих столбцов таблиц фактов и измерений куба.

Мера или атрибут (уровень)	Элементы	Исходная таблица	Исходный столбец	Образец значения столбца
Мера «Посылки»	Неприменимо	ImportsFactTable	Посылки	12
Мера «Последняя дата»	Неприменимо	ImportsFactTable	Последняя дата	03-май-99
Уровень категории «Маршрут» в измерении «Маршрут»	не наземный, наземный	RouteDimensionTable	Route_Category	Не наземный
Атрибут «Маршрут» в измерении «Маршрут»	воздушный, морской, дорожный, железнодорожный	RouteDimensionTable	Маршрут	Морской
Атрибут «Полушарие» в измерении «Источник»	Восточное полушарие, западное полушарие	SourceDimensionTable	Полушарие	Восточное полушарие

Атрибут «Континент» в измерении «Источник»	Африка, Азия, Австралия, Европа, Северная Америка, Южная Америка	SourceDimensionTable	Континент	Европа
Атрибут «Полугодие» в измерении «Время»	Первое полугодие, второе полугодие	TimeDimensionTable	Полугодие	Второе полугодие
Атрибут «Квартал» в измерении «Время»	Первый квартал, второй квартал, третий квартал, четвертый квартал	TimeDimensionTable	Квартал	Третий квартал

Данные в одной ячейке куба обычно выводятся из нескольких строк таблицы фактов. Например, ячейка куба на пересечении элемента «воздушный», элемента «Африка» и элемента «1 квартал» содержит значение, выведенное статистическим вычислением следующих рядов в таблице фактов **ImportsFactTable**.

Import_ReceiptKey	RouteKey	SourceKey	TimeKey	Посылки	Последняя дата
3516987	1	6	1	15	10-января-99
3554790	1	6	1	40	19-января-99
3572673	1	6	1	34	27-января-99
3600974	1	6	1	45	02-фев-99
3645541	1	6	1	20	09-фев-99
3674906	1	6	1	36	17-фев-99

В предыдущей таблице каждая строка содержит одни и те же значения в столбцах **RouteKey**, **SourceKey** и **TimeKey**; это означает, что эти столбцы ссылались на одну и ту же ячейку куба.

Показанный здесь пример представляет очень простой куб, в том смысле, что это куб с единственной группой мер, а все таблицы измерений соединены с таблицей фактов по схеме «звезда». Другая схема — это схема «снежинка», в которой одна или несколько таблиц измерений присоединяются к другой таблице измерения, а не напрямую к таблице фактов.

В приведенном здесь примере содержится только одна таблица фактов. Когда в кубе есть несколько таблиц фактов, меры каждой из них организуются в группы мер, причем группа мер связана с соответствующим набором измерений согласно заданным связям измерений. Эти связи определяются указанием участвующих таблиц в представлении источника данных и гранулярности связи.

Deductor Studio 5.2

Deductor — это аналитическая платформа, основа для создания законченных прикладных решений в области анализа данных. Реализованные

в Deductor технологии позволяют на базе единой архитектуры пройти все этапы построения аналитической системы: от консолидации данных до построения моделей и визуализации полученных результатов.

Аналитическая платформа Deductor состоит из пяти частей:

- Warehouse – хранилище данных, консолидирующее информацию из разных источников;
- Studio – приложение, позволяющее пройти все этапы построения прикладного решения, рабочее место аналитика;
- Viewer – рабочее место конечного пользователя, одно из средств тиражирования знаний (т.е. когда построенные аналитиком модели используют пользователи, не владеющие технологиями анализа данных);
- Server – служба, обеспечивающая удаленную аналитическую обработку данных;
- Client – клиент доступа к Deductor Server. Обеспечивает доступ к серверу из сторонних приложений и управление его работой.

После запуска главное окно Deductor Studio выглядит следующим образом (рисунок 6).

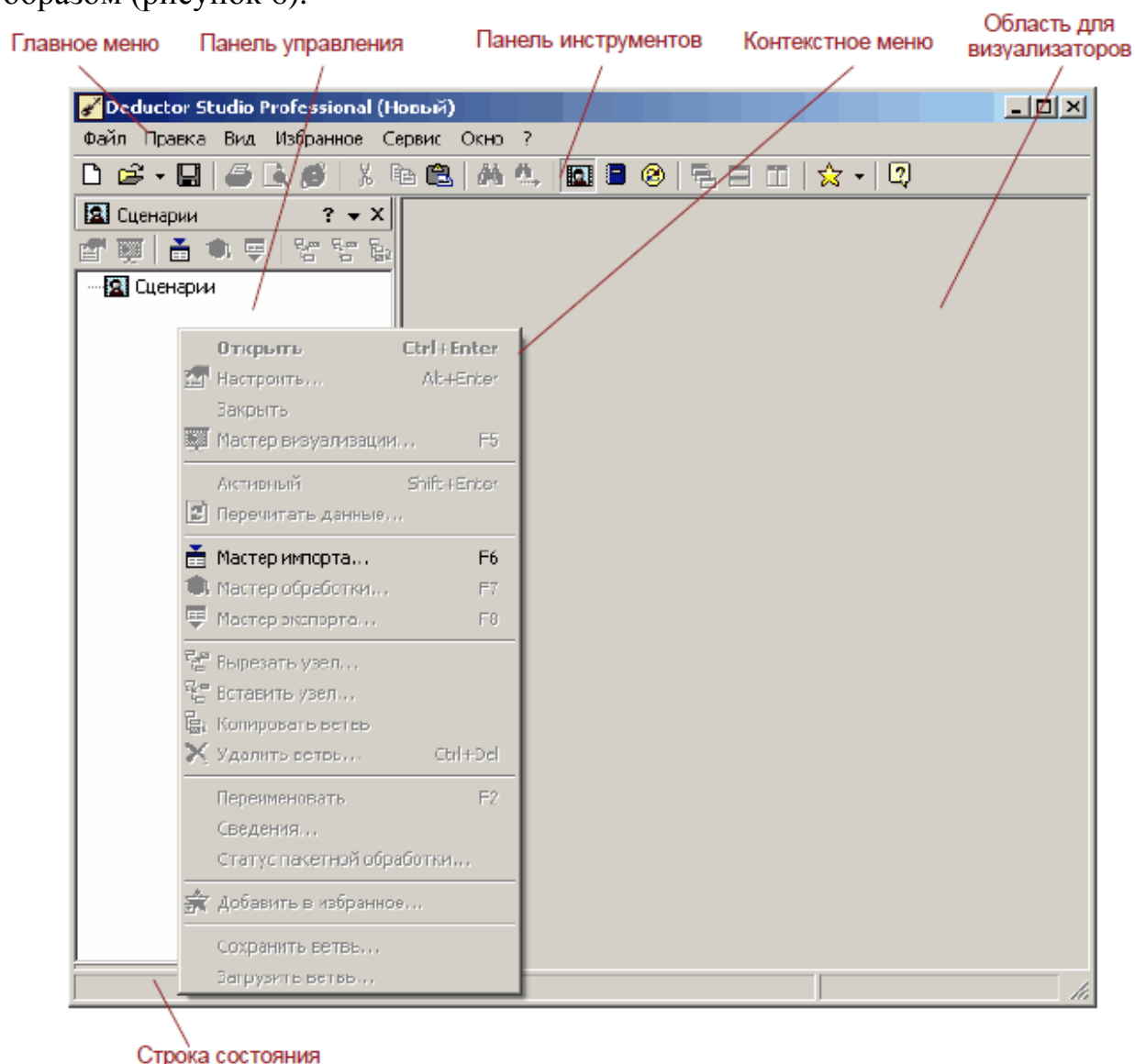




Рисунок 6 – Главное окно Deductor Studio

По умолчанию панель управления представлена одной вкладкой Сценарии. Кроме того, доступны еще две вкладки: Отчеты и Подключения. Сделать их видимыми можно следующими способами:

- главное меню Вид ► Отчеты и Вид ► Подключения;
- кнопки  и  на панели инструментов.

Можно производить «drag & drop» манипуляции с вкладками, меняя их расположение и порядок.

При нажатии правой кнопки мыши на любой вкладке появляется контекстное меню.

- Скрыть – делает вкладку невидимой;
- Переименовать – переименовывает название вкладки;
- Закладки – переключается на выбранную закладку;
- Верх/Низ – задает расположение названий вкладок: вверху либо внизу;
- Помощь – открывает раздел справки.

В Deductor Studio ключевым понятием является проект. Это файл с расширением *.ded, по структуре соответствующий стандартному xml-файлу. Он хранит в себе:

- последовательности обработки данных (сценарии);
- настроенные визуализаторы;
- переменные проекта и служебную информацию.

Создать новый проект можно следующими способами:

- главное меню Файл ► Создать;
- кнопка Создать новый проект на панели инструментов;
- клавиша Ctrl+N.

Открытие существующего проекта:

- главное меню Файл ► Открыть;
- кнопка Открыть проект на панели инструментов;
- клавиша Ctrl+O.

Открыть проект можно еще одним способом – в главном меню Файл ► История найти имя проекта. Способ работает в том случае, если вы недавно открывали этот проект, и он сохранился в менеджере историй проектов.

В одной запущенной копии Deductor Studio можно открыть только один проект.

В Deductor Studio для аналитика основополагающим понятием является сценарий.

Сценарий представляет собой последовательность операций с данными, представленную в виде иерархического дерева. В дереве каждая операция образует узел, заголовок которого содержит: имя источника данных, наименование применяемого метода обработки, используемые при этом поля и т.д. Кроме этого, слева от наименования узла стоит значок, соответствующий типу операции.

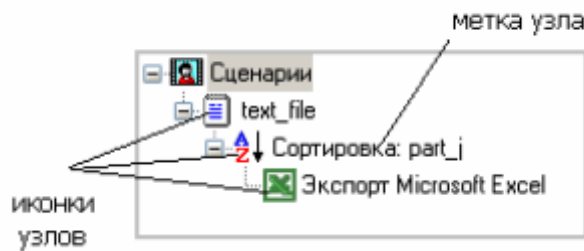



Рисунок 7 – Сценарий в Deductor Studio

Создание нового узла импорта осуществляется с помощью Мастера импорта. Вызвать мастер можно следующими способами:

- кнопка  на панели инструментов закладки Сценарии;
- клавиша F6;
- контекстное меню Мастер импорта...

При вызове мастера импорта откроется окно первого шага мастера.

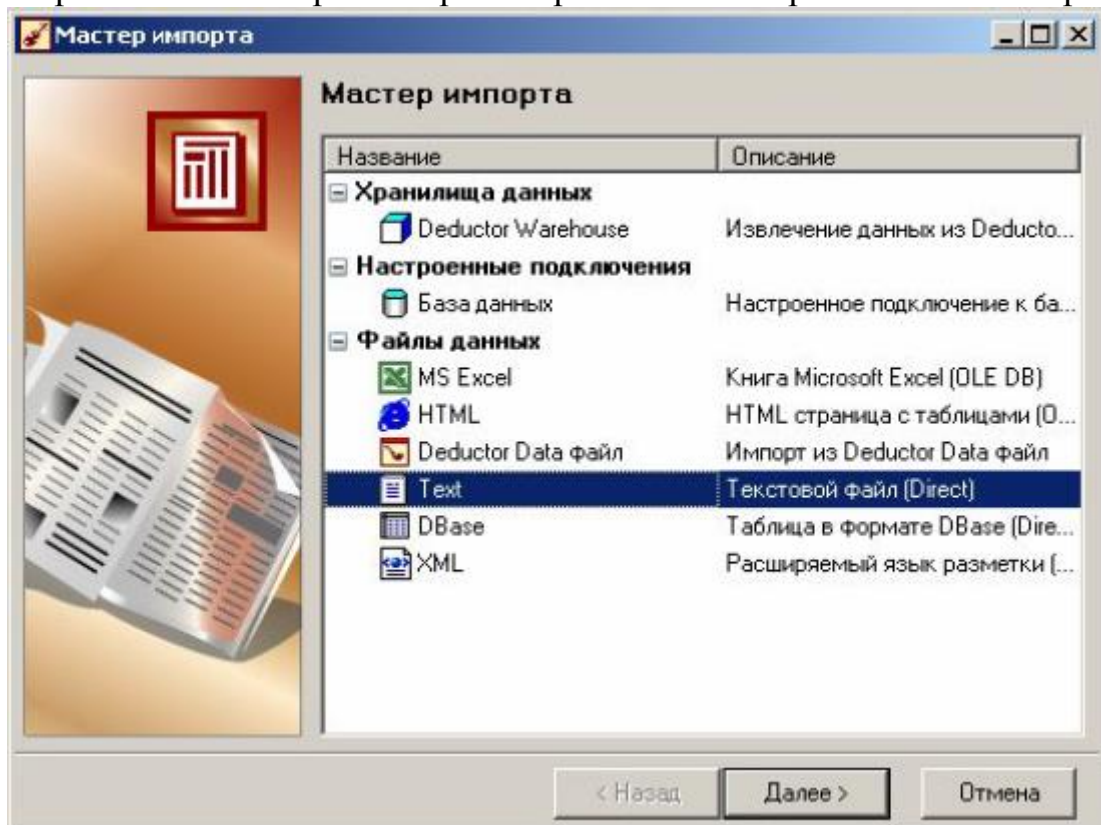



Рисунок 8 – Мастер импорта в Deductor Studio

Создание нового узла обработки осуществляется с помощью Мастера обработки. Вызвать мастер можно следующими способами:

- кнопка  на панели инструментов закладки Сценарии;
- клавиша F7;
- контекстное меню Мастер обработки...

При вызове мастера обработки откроется окно первого шага мастера.

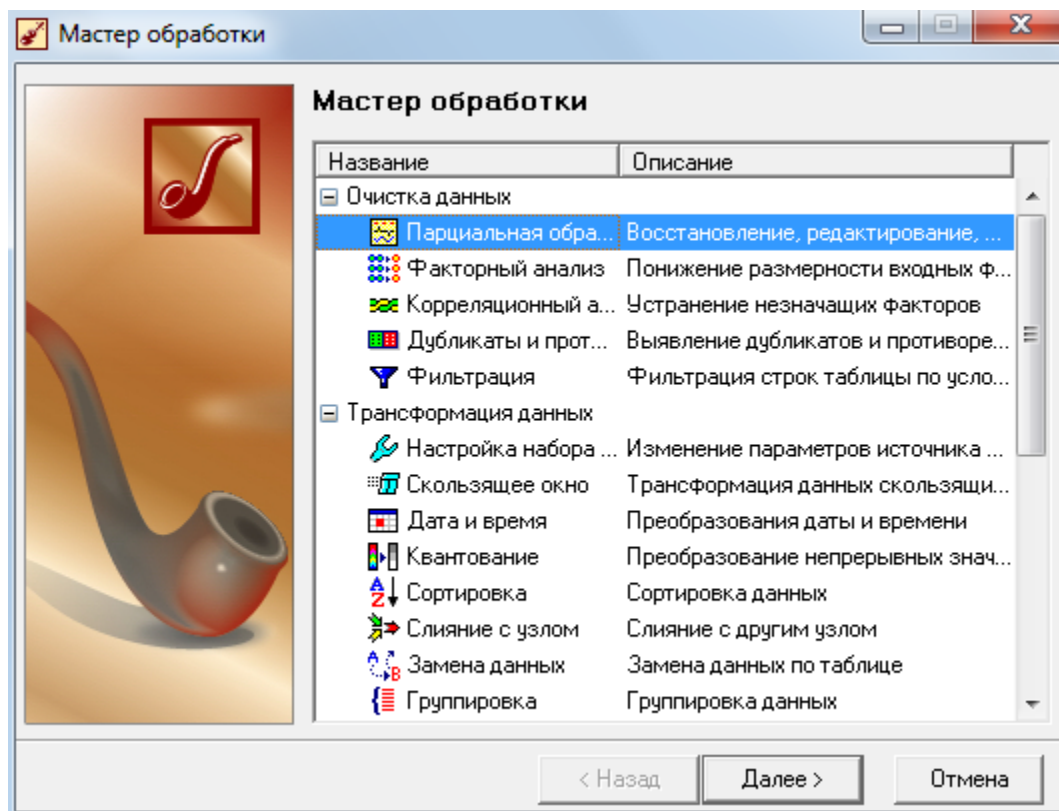



Рисунок 9 – Мастер обработки в Deductor Studio

Создание нового узла экспорта осуществляется с помощью Мастера экспорта. Вызвать мастер можно следующими способами:

- кнопка  на панели инструментов закладки Сценарии;
- клавиша F8;
- контекстное меню Мастер экспорта...

В нем все приемники данных сгруппированы по следующим 5 категориям:

- хранилища данных ;
- базы данных;
- файлы;
- Web-серверы;
- прочее.

Кроме команд вызова мастеров, к каждому узлу применимы базовые операции.


Список доступных операций.

1 Открытие узла – узел запускается на выполнение, причем выполняются все родительские узлы, а справа открываются визуализаторы, настроенные для данного узла. В интерактивном режиме для каждого узла должен быть настроен хотя бы один визуализатор, например, Таблица или Сведения. Операция вызывается:

- двойной щелчок мышью на узле;
- клавиши Ctrl+Enter;
- контекстное меню Открыть.

2 Настройка узла – вызывается мастер импорта, мастер обработки или мастер экспорта, в зависимости от типа узла, для изменения параметров обработки, производимой в узле.

Операция вызывается:

- кнопка ;
- клавиши Alt+Enter;
- контекстное меню Настроить....

3 Активация/деактивация узла – узел может быть либо активным, либо неактивным. Если узел неактивный, то, сделав его активным, выполнится сценарий для этого узла, но визуализаторы отображены не будут. Делая узел неактивным, закрываются все визуализаторы для него и для всех подчиненных узлов, а сам узел и подчиненные узлы превращаются в неактивные. Эта операция может быть использована для освобождения памяти. Операция активации/деактивации вызывается:

- клавиши Shift+Enter;
- контекстное меню Активный...

4 Перечитать данные узла – все узлы до корневого включительно будут закрыты, а затем выполнена ветка сценария от корневого до текущего узла. Операция вызывается:

- контекстное меню Перечитать данные е...

5 Вырезать узел – удаляет текущий узел из сценария обработки. Все его потомки при этом перемещаются на один уровень вверх и начинают подчиняться родителю удаленного узла.

Операция вызывается:


- кнопка ;
- контекстное меню Вырезать узел .

6 Вставить узел – вставляет перед текущим узлом сценария новый узел и вызывает для него мастер обработки. Вставить узел перед узлом импорта данных нельзя. Операция вызывается:


- кнопка ;
- контекстное меню Вставить узел .

После вставки нового узла или удаления существующего узлы-потомки могут стать неработоспособными, в зависимости от обработки, выполняемой новым узлом.

7 Копировать ветвь – копирует ветвь сценария, начиная с текущего узла и включая все его потомки. Операция вызывается:

- кнопка ;
- контекстное меню Копировать ветвь;
- при помощи механизма drag & drop – выделив узел, и, удерживая нажатой клавишу Ctrl, указать курсором мыши на новый узел, который должен стать родителем старого. При этом переносимая ветка целиком скопируется в новое место.

8 Удалить ветвь – удаляет узел сценария и все его подузлы. Удаленная ветвь восстановлению не подлежит, поэтому к данной операции необходимо подходить с осторожностью. Операция вызывается:


- кнопка ;
- клавиши Ctrl+Del;
- контекстное меню Удалить ветвь.

9 Перенос ветви – переносит ветку сценария к новому узлу. Операция производится аналогично копированию ветви с помощью drag & drop без удерживания клавиши Ctrl.

10 Переименовать – позволяет изменить метку текущего узла. Операция вызывается:

- клавиша F2;
- контекстное меню Переименовать...

11 Сведения – открывает диалоговое окно Сведения для текущего узла. В нем редактируется имя, метка и описание к узлу. Операция вызывается:

- контекстное меню Сведения ...;
- открыв скрытую панель узла с помощью кнопки  и нажать там одну из кнопок:

Имя, Метка или Описание.

Имя узла может быть задано только латинскими символами, тогда как метка – любыми. Кроме того, имя узла должно быть уникально в пределах одного сценария. Как правило, необходимости в переименовании имен узлов не возникает.

12 Статус пакетной обработки – устанавливает статус пакетной обработки для узла.

13 Добавить в Избранное – текущий узел добавляется в список избранных узлов.

14 Сохранение ветви – вызывается стандартный диалог Сохранение, в котором можно указать путь и имя файла для сохранения ветви сценария, начинающейся с текущего узла.

Операция вызывается:

- контекстное меню Сохранить ветвь.

15 Загрузка ветви – вызывает стандартный диалог Открытие файла, в котором можно указать путь и имя файла, хранящего ветвь сценария. Загруженная ветвь сценария станет потомком текущего узла. Ветвь, начинающаяся с узла импорта данных, будет добавлена в проект как новая корневая ветвь. Операция вызывается:

- контекстное меню Загрузить ветвь.

По умолчанию ветвь сценария имеет расширение *.deb.

Структурированный текстовый файл с разделителями – один из самых распространенных форматов хранения данных. Такой файл представляет собой обычный текстовый файл, столбцы данных в котором разделены однотипными символами-разделителями, например символами табуляции, пробела, точки с запятой и т.д.

Процесс импорта данных из текстового с разделителями файла в мастере импорта (категория Т екстовой файл (Direct)) содержит следующие шаги:

- указание имени файла;
- настройка параметров импорта;
- настройка импортируемых полей;
- запуск процесса импорта;
- выбор способа визуализации;
- задание сведений об узле.

На шаге Указание имени файла, нажав кнопку , необходимо выбрать имя текстового файла (расширения *.txt, *.csv), из которого следует выполнить импорт данных. После этого в поле «Имя файла» окна Мастера импорта появится имя выбранного файла и путь.

Допускается вручную ввести путь к файлу в строке поля Имя файла.

Имеется возможность использовать как абсолютные, так и относительные пути для файлов.

Они указываются относительно текущей директории Deductor. При открытии Deductor текущей директорией является директория файла проекта. Поэтому, если файл проекта и текстовые файлы располагаются в одной папке, то использование относительных путей в Мастере импорта позволит не перенастраивать узлы импорта при изменении расположения папки на жестком диске.

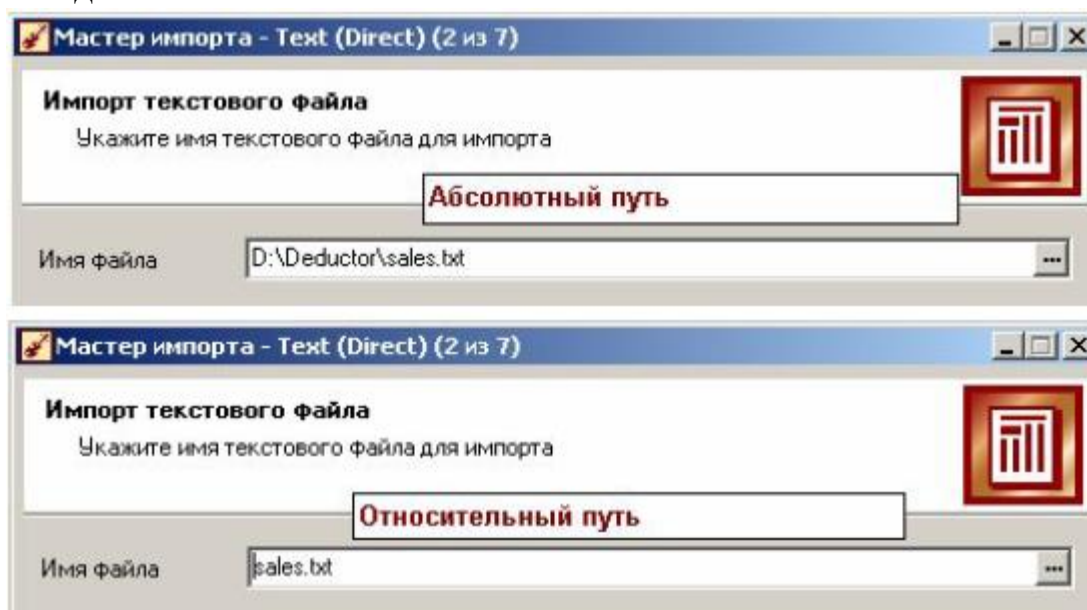


Рисунок 10 – Мастер импорта в Deductor Studio

Здесь также доступны настройки:

- начать импорт со строки – номер строки, начиная с которой будет делаться импорт данных из файла;
- флаг Первая строка является заголовком – установка флажка означает, что узел будет импортировать данные с учетом того, что все записи первой строки являются заголовками столбцов;

– кодировка – ANSI (Windows) или ANCI (MS DOS).

На шаге Настройка параметров импорта нужно настроить параметры импорта данных из текстового файла, так как существует несколько форматов структурированных текстовых файлов. Доступные опции:

– переключатель Формат исходных данных, который определяет символ-разделитель в файле (например: символ табуляции, пробел, запятая). Разделитель чаще всего присутствует. Если же нет, то нужно выбрать переключатель Фиксированной ширины (поля имеют заданную ширину), а позже установить ширину каждого поля.

– Ограничитель с трок – при задании данного параметра необходимо указать, какой именно ограничитель строкового значения нужно использовать при импорте данных из текстового файла. Обычно таким ограничителем является символ двойной кавычки ".

– Разделитель дробной и целой части числа – при задании данного параметра необходимо указать символ, разделяющий дробную и целую части в числовых значениях, содержащихся в файле.

– Разделитель компонентов даты – указывается символ, разделяющий компоненты даты в соответствующих значениях, содержащихся в файле.

– Разделитель компонентов времени – указывается символ, разделяющий компоненты времени в соответствующих значениях, содержащихся в файле.

– Форматы Даты/Времени – указываются форматы даты/времени, используемые в импортируемом файле.

– Представление значений – опция для полей логического типа, которое может принимать одно из трех значений – истина (true), ложь (false) и пустое значение (null).

Определяет регламент записи в эти значения. Так, при настройках по умолчанию для любого логического поля значение Да будет восприниматься как истина, Нет – как ложь.

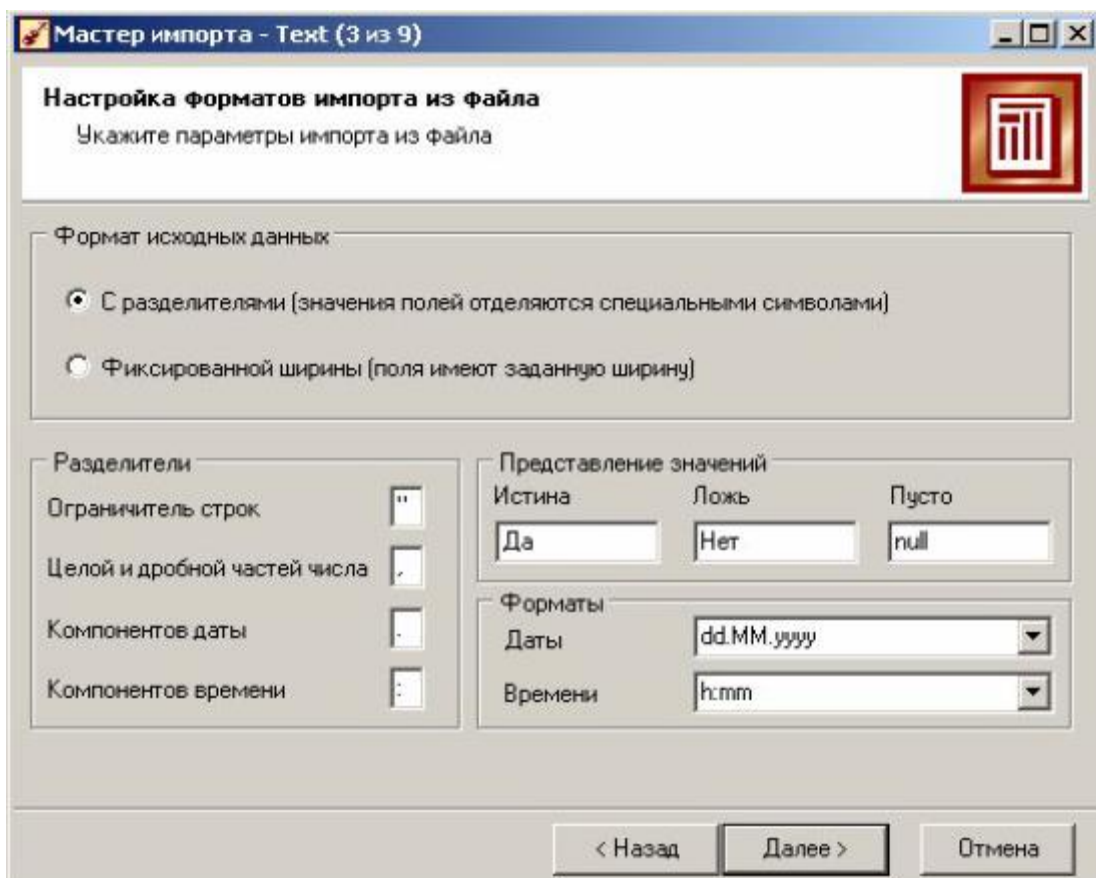


Рисунок 11 – Мастер импорта (настройка форматов импорта из файлов)

Следующее окно мастера зависит от установленного переключателя в флажке **Формат исходных данных**. Если был выбран формат **С разделителями**, то появится вкладка, на которой нужно явно указать символ-разделитель (по умолчанию – табуляция). Здесь же находится флаг **Считать последовательные разделители одним** – в случае последовательно идущих символов-разделителей они будут восприниматься за один. Такое бывает, например, когда символом-разделителем выступают несколько пробелов.

Предпросмотр текстового файла в виде таблицы внизу (загружаются только первые 10 строк) позволяет убедиться в корректности выбора настроек импорта даже не запуская его.

Мастер импорта - Text (4 из 9)

Параметры импорта файла с разделителями
 Укажите символ-разделитель столбцов и другие вспомогательные параметры импорта

Символом-разделителем является

☒ Символ табуляции
 ☐ Пробел
 ☐ Точка
 ☐ Точка с запятой
 ☐ Запятая
 ☐ Другой

☐ Считать последовательные разделители одним

Банк	Активы			Структура п...	
	Вложения в векселя, %	Кредиты частным лицам, %	Кредиты предприятиям и организациям, %	Собственный капитал, %	Привлеченный МБК, %
Сбербанк	0	23	56	10	3
ВТБ	2	1	46	17	31
Газпромбанк	1	3	40	12	14

< Назад Далее > Отмена

Рисунок 12 – Мастер импорта (параметры импорта файла с разделителями)

Если был выбран флаг формат Фиксированной ширины, то появится вкладка, на которой нужно задать границы каждого поля. Создание, как и удаление маркера границы производится одним щелчком мыши. Двигая маркеры границ столбцов, можно изменять их, если они расставлены неправильно. Данные, распределенные по столбцам, показываются в области предварительного просмотра.

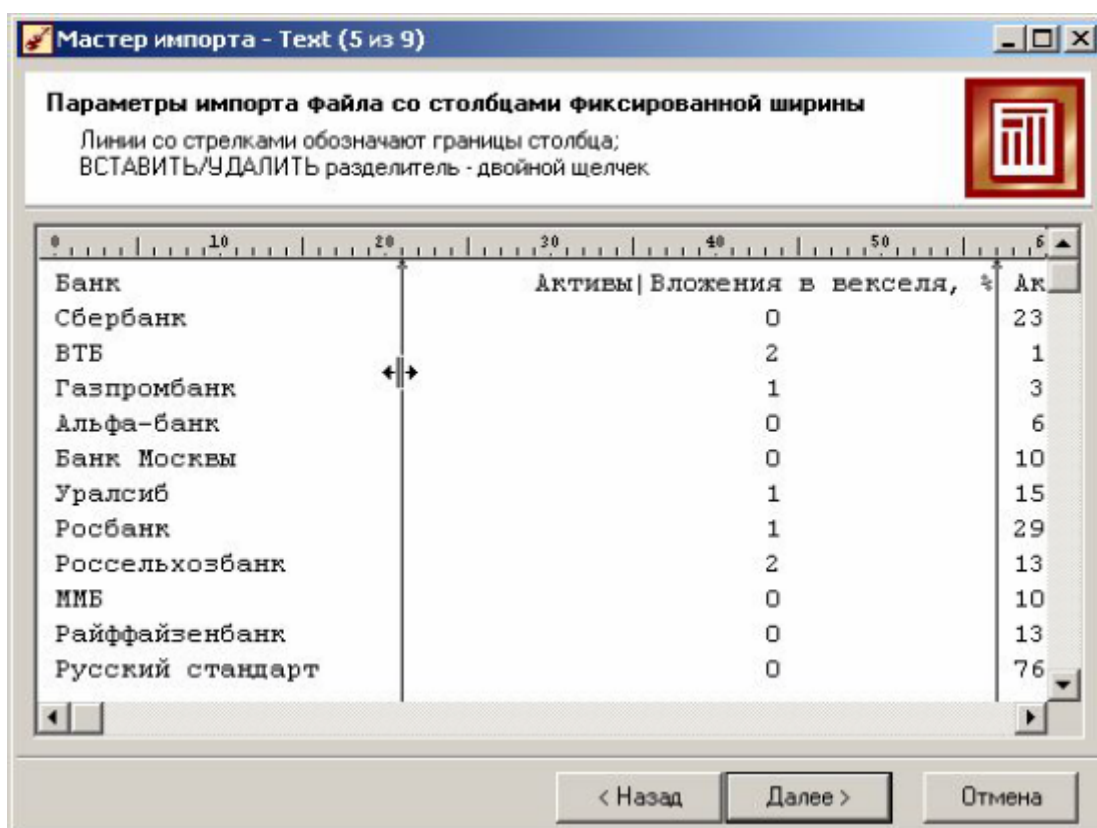


Рисунок 13 – Мастер импорта (параметры импорта файла со столбцами фиксированной ширины)

На шаге Настройка параметров столбцов нужно настроить следующие параметры столбцов импортируемых данных, указав соответствующие значения в полях.

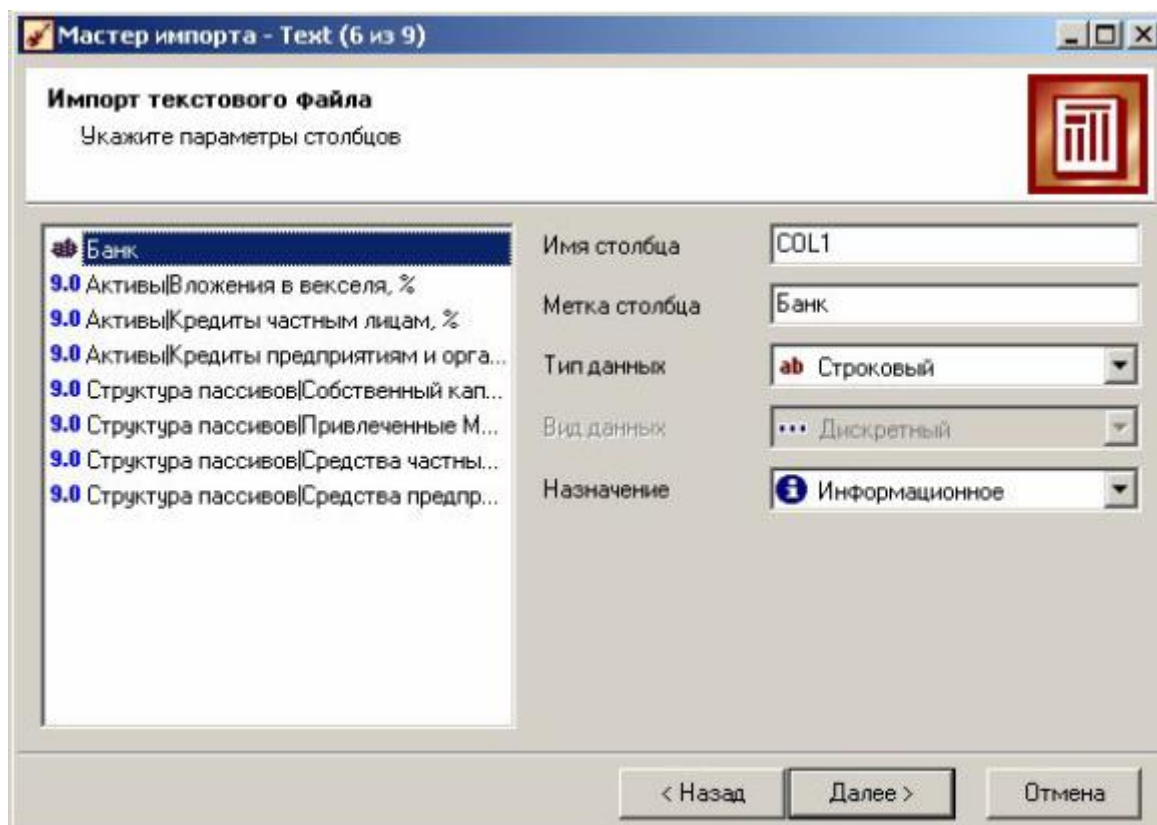


Рисунок 14 – Мастер импорта (импорт текстового файла)

Обработчик Настройка набора данных позволяет:

- изменить имя, метку, тип, вид и назначение полей текущего набора данных;
- изменить порядок следования столбцов в наборе данных;
- скрыть столбцы набора данных;
- задать опцию кэширования выходного набора.

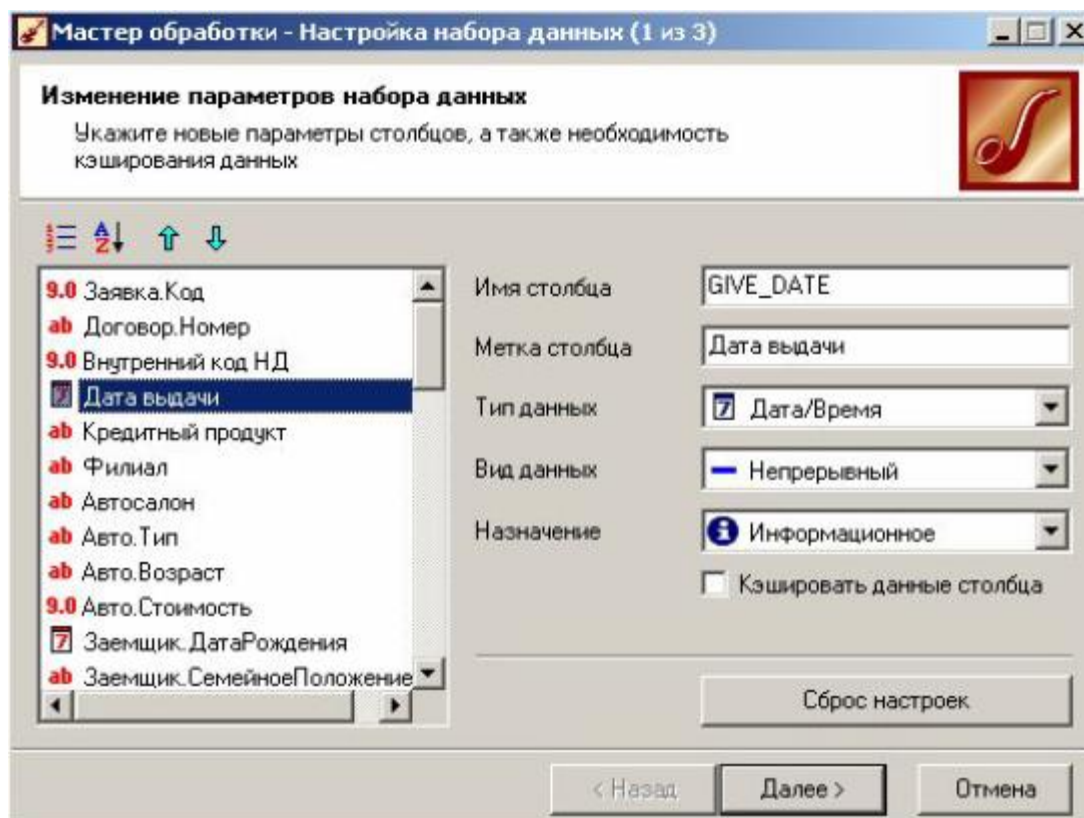





Рисунок 15 – Мастер обработки (измерение параметров набора данных)

Изменение имени или метки поля удобно в тех случаях, когда имена столбцов могут измениться в источнике данных или при перенастройке узлов верхних уровней. В этом случае в узле Настройка набора данных имя исходного столбца заменяется другим, на которое и настраиваются все дочерние узлы. После такой операции изменение имен полей на верхних уровнях не потребует перенастройки всех дочерних узлов в дереве сценариев.

Тип, вид и назначение можно изменить у нескольких столбцов одной операцией. Для этого достаточно их выделить, удерживая нажатой клавишу Ctrl или Shift.

Если параметры столбца были изменены, цвет иконки столбца меняется на красный. Для установки первоначальных параметров столбцов необходимо выделить столбец или список столбцов и нажать на кнопку Сброс параметров.

Чтобы скрыть столбец из набора данных, нужно задать ему назначение  Неиспользуемое.

Изменить порядок следования столбцов в наборе данных можно при помощи клавиш  .

Кэширование – это загрузка часто используемой информации в оперативную память для быстрого доступа к ней, минуя многократные считывания с жесткого диска. Кэширование может заметно повысить

скорость работы сценария в ряде случаев (использование кэширования не входит в базовые навыки работы с Deductor).

К каждому узлу сценария, который содержит структурированный набор данных, всегда предлагается несколько визуализаторов. Мастер визуализации в интерактивном пошаговом режиме позволяет выбрать и настроить наиболее удобный способ представления данных. В зависимости от выбранного способа будут настраиваться различные параметры, а Мастер, соответственно, будет содержать различное число шагов. Первый шаг мастера визуализации будет одинаков для всех видов, поскольку на нем и производится выбор визуализатора.

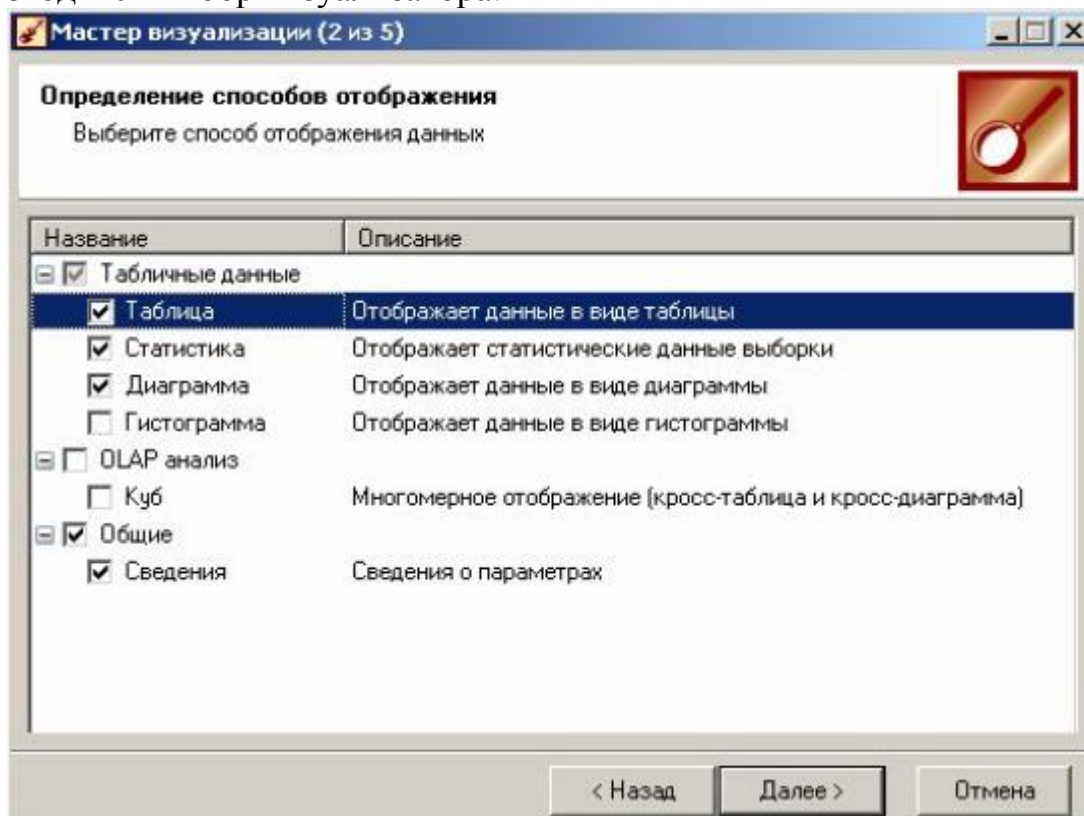



Рисунок 16 – Мастер визуализации

Вызов мастера визуализации:

- кнопка  на панели инструментов закладки Сценарии;
- клавиша F5;
- контекстное меню Мастер визуализации...

Мастер визуализации запускается для выделенного узла сценария. Кроме того, этот мастер всегда является продолжением мастера обработки, т.е. активизируется при создании (настройке) любого узла.

Желаемые способы отображения следует пометить флажками. Одновременно может быть выбрано несколько визуализаторов, при этом каждый из них будет открыт в отдельном окне.

Базовыми визуализаторами в Deductor являются следующие:

- Таблица;

- Статистика;
- Сведения.

Дата кредитования	Сумма кредита	Срок кредита	Цель кредитования	Частная собственность
01.01.2003	7000	6	Иное	<input type="checkbox"/>
01.01.2003	7500	6	Иное	<input checked="" type="checkbox"/>
01.01.2003	14500	12	Покупка товара	<input checked="" type="checkbox"/>
01.01.2003	15000	6	Покупка товара	<input type="checkbox"/>
01.01.2003	32000	12	Иное	<input checked="" type="checkbox"/>
01.01.2003	11500	6	Турпоездки, развлечения и т.п.	<input type="checkbox"/>
01.01.2003	5000	6	Покупка и ремонт недвижимости	<input checked="" type="checkbox"/>
01.01.2003	61500	30	Покупка товара	<input checked="" type="checkbox"/>
01.01.2003	13500	12	Оплата услуг (мед., юрид. и т.п.)	<input type="checkbox"/>
01.01.2003	25000	18	Покупка товара	<input checked="" type="checkbox"/>
01.01.2003	25500	24	Покупка товара	<input checked="" type="checkbox"/>
01.01.2003	9500	6	Покупка товара	<input checked="" type="checkbox"/>
01.01.2003	53000	24	Иное	<input checked="" type="checkbox"/>
02.01.2003	27500	18	Покупка товара	<input checked="" type="checkbox"/>
02.01.2003	4000	6	Оплата услуг (мед., юрид. и т.п.)	<input type="checkbox"/>
02.01.2003	40500	24	Покупка и ремонт недвижимости	<input checked="" type="checkbox"/>
02.01.2003	51500	36	Покупка и ремонт недвижимости	<input type="checkbox"/>
02.01.2003	7000	6	Оплата услуг (мед., юрид. и т.п.)	<input type="checkbox"/>
02.01.2003	8500	6	Турпоездки, развлечения и т.п.	<input type="checkbox"/>
02.01.2003	23500	12	Иное	<input checked="" type="checkbox"/>
02.01.2003	16500	12	Покупка товара	<input type="checkbox"/>
02.01.2003	46500	36	Покупка товара	<input checked="" type="checkbox"/>
02.01.2003	58000	48	Покупка и ремонт недвижимости	<input type="checkbox"/>
02.01.2003	58500	42	Покупка товара	<input type="checkbox"/>
02.01.2003	20500	12	Покупка товара	<input checked="" type="checkbox"/>
02.01.2003	3500	6	Покупка и ремонт недвижимости	<input checked="" type="checkbox"/>
03.01.2003	27500	12	Покупка и ремонт недвижимости	<input checked="" type="checkbox"/>


Рисунок 17 – Визуализатор «Таблица»

В таблице каждое поле набора данных размещается в отдельном столбце. Столбцы озаглавлены метками полей, а если метка не была задана, то именами полей. Ширину и порядок столбцов можно менять при помощи мыши.

В таблице можно настроить объединение заголовков столбцов. Например, есть два заголовка Продажи Сумма и Продажи Количество. Если переименовать (например, с помощью обработчика Настройка набора данных) метку первого столбца в Продажи|Сумма, а второй – в Продажи|Количество, то получим объединение заголовка в шапке таблицы.

Продажи	
Сумма	Количество

Однажды настроенный вид таблицы (к примеру, с различными фильтрами, форматами и видимостью столбцов и т.п.) можно сохранить, чтобы впоследствии быстро вернуться к нему.

Для этого в раскрывающемся по кнопке  списке нужно выбрать пункт Сохранить конфигурацию... и далее ввести ее название. Загрузить новую конфигурацию, можно, выбрав ее из списка конфигураций.

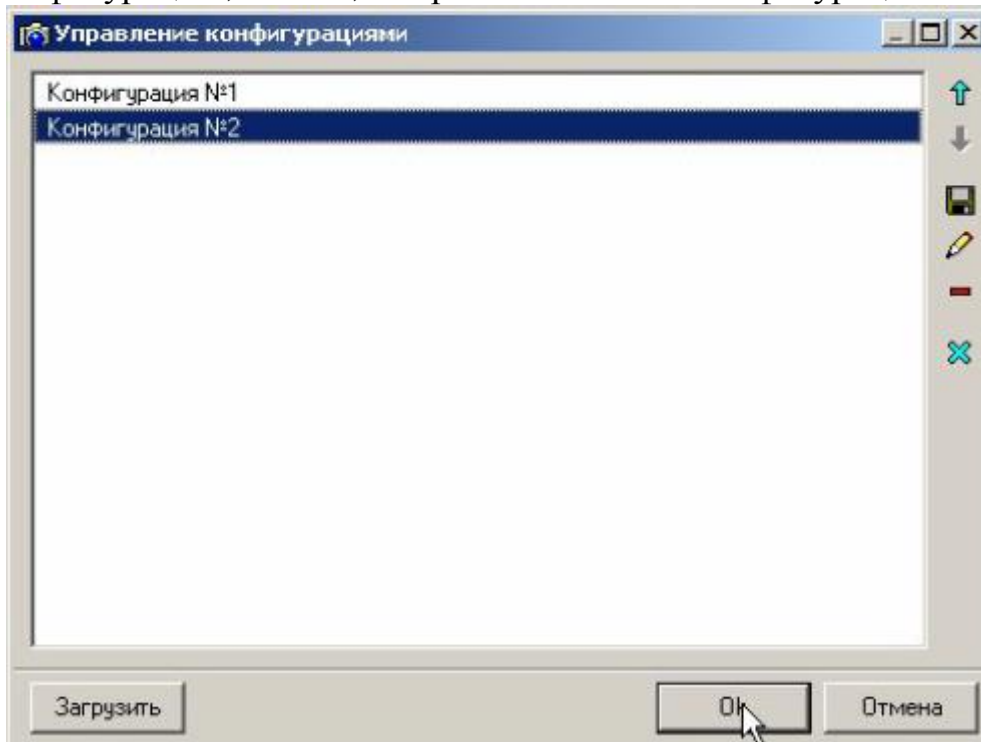


Рисунок 18 – Управление конфигурациями

При вызове настройки полей появляется соответствующее диалоговое окно. В нем можно скрыть или сделать видимыми различные поля таблицы, определить способ выравнивания содержимого, ширину поля, а также задать формат отображения числовых данных и дат.

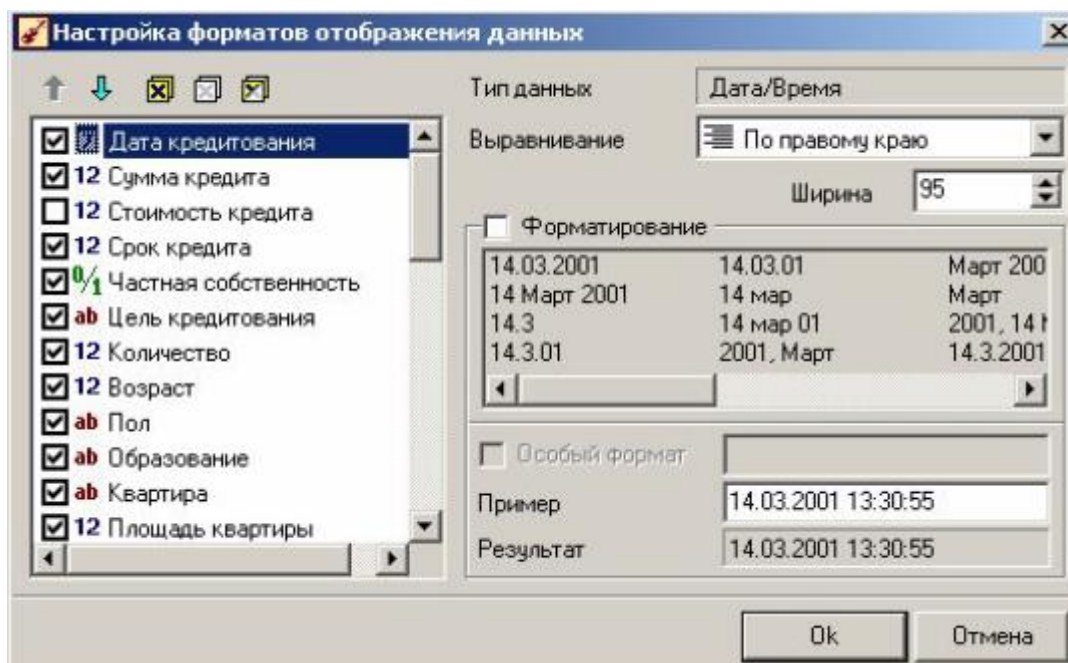





Рисунок 19 – Настройка форматов отображения данных


Кнопка  переключает способ отображения набора данных, который может быть не только табличным, но и в виде формы. Это удобно, когда набор данных содержит большое количество столбцов.

Кнопка  открывает окно настройки условий фильтрации на набор данных.

При включенном фильтре цвет кнопки меняется на , а цвет заголовков столбцов, которые участвуют в фильтре, изменяется на красный:

Дата (Год + Месяц)	Количество
2002-M01	355000
2002-M02	340000
2002-M03	405000
2002-M04	452000
2002-M05	464000
2002-M06	437000

Рисунок 20 – Использование фильтра

Кнопка  открывает визуализатор Статистика, но не в отдельном вкладке, а в нижней части визуализатора Таблица.

Статистик а служит для отображения основных статистических характеристик набора данных конкретного узла.

Статистические характеристики отображаются в таблице по каждому полю выборки. В верхней части окна статистики отображается общее количество записей в наборе данных. Панель инструментов окна статистики позволяет управлять отображением статистических характеристик (среднее,

минимум, максимум и т.п.) с помощью группы кнопок


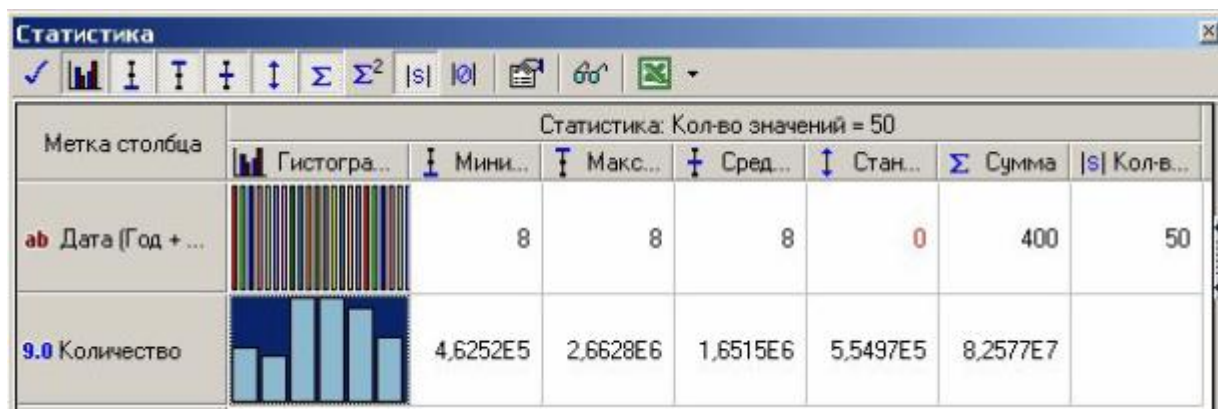



Рисунок 21 – Визуализатор «Статистика»

Для полей дискретного типа, кроме прочих, всегда рассчитываются следующие статистические показатели:

- количество уникальных значений,
- количество пустых значений.

Просмотреть список уникальных значений можно следующими способами:

- двойной щелчок по ячейке Количество уникальных значений или по ячейке Гистограмма,
- кнопка Обзор статистики .

Для поля непрерывного типа в обзоре статистики строится гистограмма распределения частот, она же в уменьшенном виде всегда показывается в соответствующем столбце.

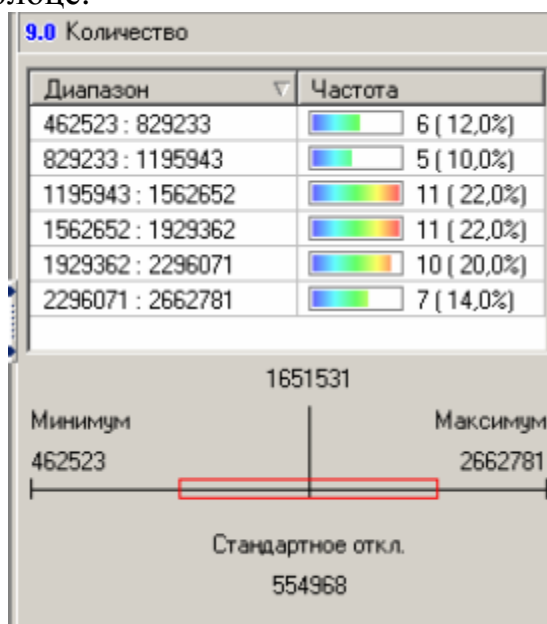


Рисунок 22 – Гистограмма распределения частот

Визуализатор Сведения позволяет просмотреть все параметры, с которыми был выполнен тот или иной процесс преобразования данных, в результате которого была сформирована новая выборка: импорт, обработка одним из методов или экспорт. Такими параметрами являются время и длительность выполняемого процесса, условия остановки, наличие первичного ключа, ограничители столбцов, разделители целой и дробной частей чисел, элементов даты и т.д.

Предусмотрено два вида представления описания: в виде дерева и текстовый. По умолчанию устанавливается вид дерева.

[-] Узел	
... Имя	146
... Метка	Данные по продажам
... Описание	
[-] Объект	Текстовый файл (..\Samples\TradeSales.txt)
... Максимальное время выполнения	0
... Время выполнения (мс)	31
... Начало процесса	2007.09.03 11:31:45
... Конец процесса	2007.09.03 11:31:46
... Время выполнения	0:00:00
... Процесс остановлен по условию останова	False
... Процесс остановлен пользователем	False
... Текстовый файл	..\Samples\TradeSales.txt
... Добавить первичный ключ	False
... Разделитель столбцов	Табуляция
... Ограничитель строк	""
... Считать последовательные разделители одним	False

Рисунок 23 – Визуализатор «Сведения»

Визуализатор в основном предназначен для оперативного анализа текущих настроек узлов и для поиска возможных ошибок.

Визуализатор Сведения является единственно доступным для узлов экспорта.

Обработчик Сортировка предназначен для изменения порядка следования записей в наборе данных в соответствии с выбранным типом сортировки.

Результатом выполнения сортировки является новый набор данных, записи в котором следуют в соответствие с заданными параметрами сортировки.

Если сортировка производится по одному полю, то все записи исходного набора данных располагаются в порядке возрастания или убывания его значений. Если сортировка производится по двум или более полям, то действуют следующие правила:

- 1 Сначала записи сортируются в заданном порядке для первого поля.
- 2 В каждом наборе одинаковых значений первого поля записи располагаются в заданном порядке для второго поля.

И так далее для всех полей, подлежащих сортировке.

Обработчик Сортировка находится в группе узлов Трансформация данных мастера обработки.

В единственном окне настройки параметров сортировки мастера обработки представлен список условий сортировки, в котором содержатся две графы:

- Имя поля – содержит имена полей, по которым следует выполнить сортировку.
- Порядок сортировки – содержит порядок сортировки данных в соответствующем поле – по возрастанию или по убыванию.

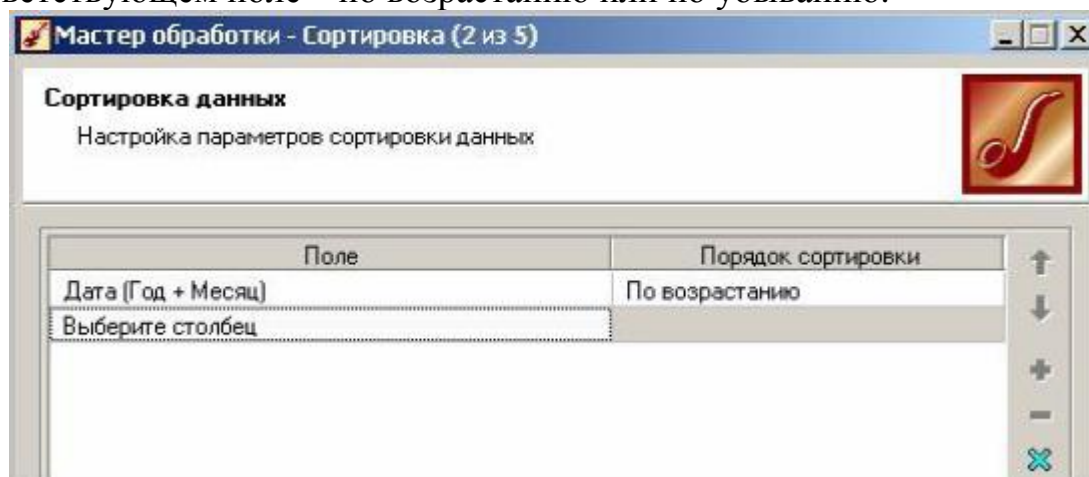


Рисунок 24 – Мастер обработки – Сортировка



Обработчик Замена данных предназначен для замены значений набора данных по таблице подстановок, которая содержит пары, состоящие из исходного значения и результирующего значения.

Пример таблицы подстановок.

Значение	Заменять на
Мск	Москва
Спб	Санкт-Петербург
Ектб	Екатеринбург

Для каждого значения исходного набора данных ищется соответствие среди исходных значений таблицы подстановки. Если соответствие найдено, то значение меняется на соответствующее выходное значение из таблицы подстановки. Если значение не найдено в таблице, оно может быть либо заменено значением, указанным для замены «по умолчанию», либо оставлено без изменений (если такое значение не указано).

В результате замены для каждого поля, которое в нем участвует, создается новое поле с префиксом `_REPLACE` как к имени, так и к метке поля. Например, для поля Город после узла Замен а данных появится новое поле `Город_REPLACE`.

Обработчик Замена данных находится в группе узлов Трансформация данных мастера обработки. В окне настройки параметров замены для каждого поля можно ввести таблицу подстановок. Добавление новой строки в таблицу подстановок производится нажатием кнопки , удаление существующей – .

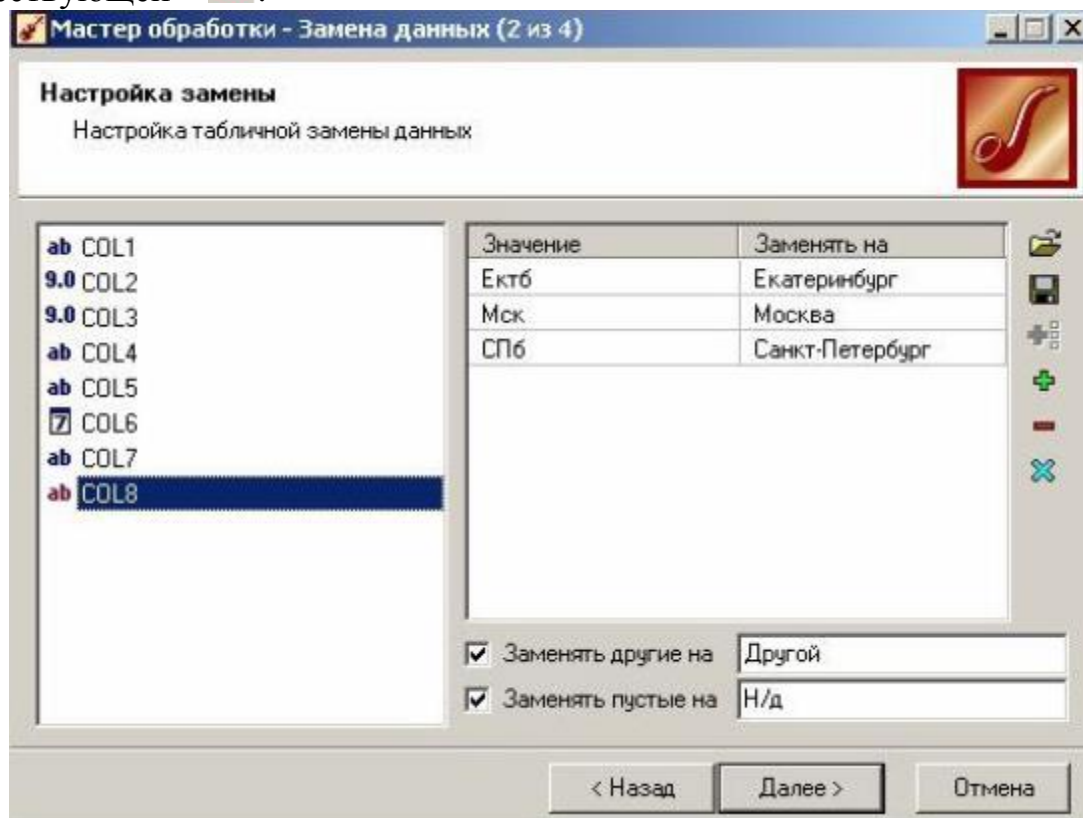



Рисунок 25 – Мастер обработки – Замена данных

В таблице подстановок должны быть заполнены два поля:

– Значение – заменяемое значение поля исходной таблицы. Если поле дискретное, то для ввода значения можно воспользоваться кнопкой выбора , где флажками отметить нужные значения. При этом откроется диалоговое окно:

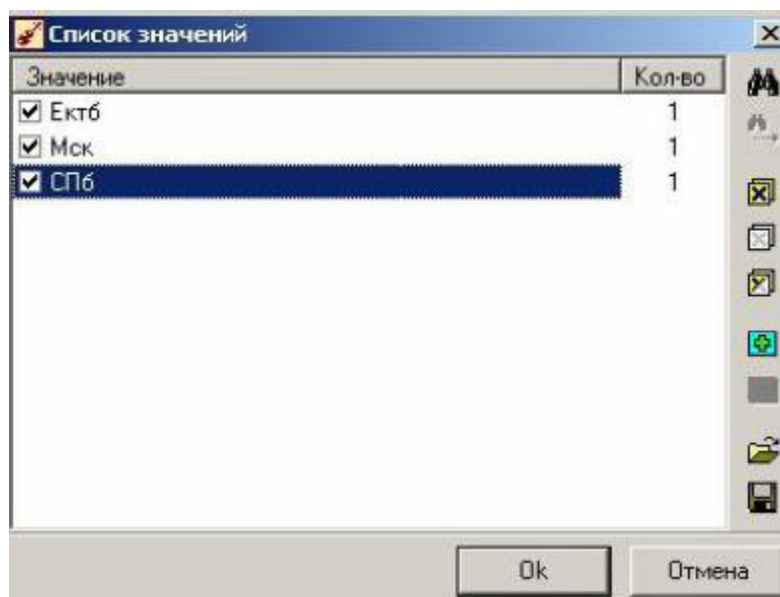



Рисунок 26 – Список значений


– Заменять на – значение для замены того, что указано в поле Значение.

Внизу таблицы подстановок расположены еще два флага, которые при необходимости можно включить:

– Заменять другие на – на какое значение следует заменить значения, не указанные в таблице замены. Для этого установите флажок и в поле напротив введите значение для замены.

– Заменять пустые на – на какое значение заменять пустые значения поля.

Таблицу подстановки, кроме непосредственного ввода, можно заполнить, загрузив ее из текстового файла (кнопка ). Формат текстового файла должен быть следующим: <заменяемое значение><символ табуляции><значение для замены>

И наоборот, список подстановок можно сохранить в текстовый файл (кнопка ).

Если по полю настроена таблица подстановок, иконка типа данных меняет свой цвет на красный. Попытка сделать замену данных с числового типа на строковый может потерпеть неудачу с выдачей соответствующего сообщения. Например, заменить все пустые значения на «н/д» в вещественном поле Сумма не получится, т.к. поле уже становится не вещественным, а строковым. Поэтому предварительно необходимо преобразовать поле Сумма в строковый тип при помощи обработчика Настройка набора данных.

Обработчик Фильтрация находится в группе узлов Очистка данных мастера обработки.


Параметры фильтрации задаются в виде списка условий, который содержит следующие столбцы.

1 Операция – позволяет установить функцию отношения «И» или «ИЛИ» между полями, для каждого из которых выполняется фильтрация. Возможна фильтрация по нескольким условиям для нескольких полей одновременно. В результате фильтрации по каждому из полей или условий будет получено отдельное множество значений. Функция в поле Операция устанавливает отношение между этими множествами. Если используется отношение «И», то в результирующий набор будут включены записи, удовлетворяющие условиям фильтрации по обоим полям. Если используется отношение «ИЛИ», то в выходной набор будут включены данные, удовлетворяющие хотя бы одному из условий.

Установка отношений возможна, только если настроены два или более условия фильтрации. Для выбора операции следует дважды щелкнуть левой кнопкой мыши в столбце Операция для соответствующего условия и из списка, открываемого кнопкой, выбрать нужную функцию отношения. По умолчанию устанавливается отношение «И».

2 Имя поля – позволяет выбрать поле, по значениям которого должна быть выполнена фильтрация. Одно и то же поле может быть использовано в нескольких условиях.

3 Условие – указывается условие, по которому нужно выполнить фильтрацию для данного поля.

Для выбора условия достаточно дважды щелкнуть мышью в соответствующей ячейке и в списке условий, открываемом кнопкой , выделить нужное условие. Доступны следующие условия фильтрации:

- (равно), < (меньше), <= (меньше или равно), > (больше), >= (больше или равно), <> (не равно) – отбираются только те записи, значения которых в данном поле удовлетворяют заданному выражению;

- пустой – отбираются только те записи, для которых в данном поле содержится пустое значение. В этом случае поле Значение не используется;

- не пустой – отбираются только те записи, для которых в данном поле не содержится пустое значение. В этом случае поле Значение не используется;

- содержит – отображаются только те записи, которые в данном столбце содержат указанное значение;

- не содержит – отображаются только те записи, которые в данном столбце не содержат указанное значение;

- в интервале, вне интервала – для числовых полей и полей типа Дата/время отбираются только те записи, значения которых в данном столбце лежат в выбранном диапазоне (вне выбранного диапазона);

- в списке, вне списка – отбираются только те записи, которые содержатся в выбранном списке (вне выбранного списка);

- начинается на, не начинается на – для строковых полей отбираются записи, значения которых в данном столбце начинаются (не начинаются) на введенную последовательность символов.

– заканчивается на, не заканчивается на – для строковых полей отбираются записи, значения которых в данном столбце заканчиваются (не заканчиваются) на введенную последовательность символов.

– первый, не первый – для полей типа Дата/время – по данному полю отбираются первые (не первые) N периодов от выбранной даты. Периодом может быть день, неделя, месяц, квартал, год. Например, если выбрать условие «первые 3 дня от 29.11.2004», то будут отобраны записи, в которых значение данного поля равно «29.11.2004», «30.11.2004», «01.12.2004» – 3 последующих дня.


– последний, не последний – для полей типа Дата/время отбираются последние (не последние) N периодов от выбранной даты. Периодом может быть день, неделя, месяц, квартал, год. Например, если выбрать условие «последние 3 дня от 29.11.2004», то будут отобраны записи, в которых значение данного поля равно «29.11.2004», «28.11.2004», «27.11.2004» – 3 предыдущих дня.

4 Значение – указывается значение(я), по которому будет производиться фильтрация записей в соответствии с заданным условием. Способ ввода значения будет различным в зависимости от типа данных и выбранного условия. Допустим, в качестве условия выбрана операция отношения «=», «<», «>» и т.д. Если данные в поле являются непрерывными (т.е. числовыми), то достаточно дважды щелкнуть мышью в соответствующей ячейке, чтобы появился курсор, затем ввести значение (число). Если поле, по которому выполняется фильтрация, имеет тип «строка» (т.е. является дискретным), то в результате двойного щелчка в столбце Значение появится кнопка выбора, которая откроет окно «Список уникальных значений», где будут отображены все уникальные значения поля и их количество. Чтобы выбрать значение для условия отбора, достаточно выделить его и щелкнуть Ok, либо просто дважды щелкнуть мышкой на нужном значении. Если выбрано условие между или не между, тогда после щелчка мышки откроется окно, в котором необходимо указать верхнюю и нижнюю границы интервала, и так далее.

Флажок Учитывать регистр учитывает регистр символов при отборе.

Внизу окна настроек в автоматическом режиме формируется выражение для фильтрации, полученное объединением всех условий, например: ([Размер ссуды, руб] в интервале [2000..5000]) И ([Цель ссуды] = 'Покупка товара').

Иногда возникает необходимость построить фильтр для дискретного поля с условием в списке, вне списка для значений, которые не существуют в наборе данных (но предполагается, что они могут появиться в будущем).

Выходом служит кнопка  Добавить значение в окне выбора списка значений. Количество записей такого «несуществующего» списочного значения всегда будет равно нулю, а строка – подкрашена светло-желтым цветом.

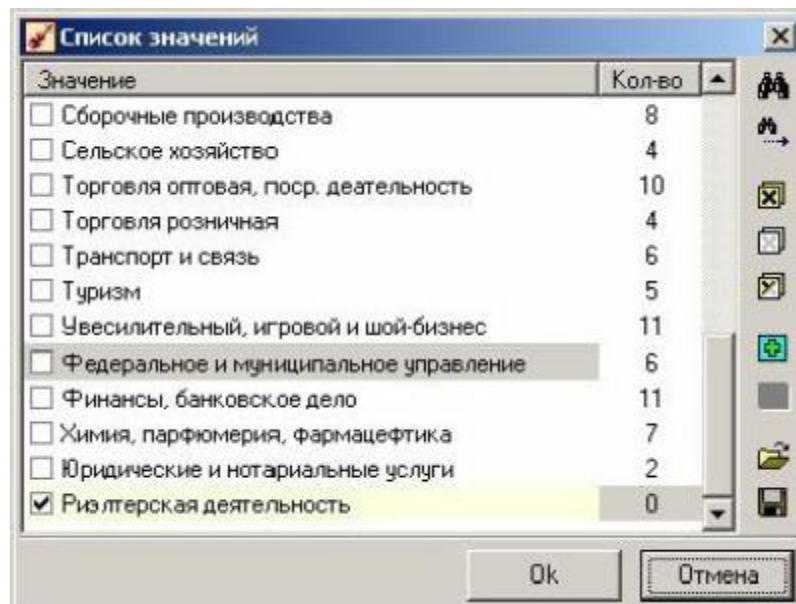


Рисунок 27 – Список значений

Задание для самостоятельной работы:

1. Создайте новый проект и сохраните его под именем test.ded. Не используйте упакованный формат файла.
2. Создайте и сохраните в любом текстовом редакторе файл следующего вида:
a,1,4.5,b,c,26/04/2007,d
a1,0,5,b1,c1,,d1
3. Импортируйте его в Deductor, корректно настроив параметры импорта. Используйте относительный путь для файла. Метку узла переименуйте в Пример импорта файла. В комментарии к узлу впишите: Текстовый файл с разделителями-запятыми.
4. Добавьте к узлу узел Настройка набора данных и задайте следующие метки к столбцам: Поле1, Поле2, Поле3 и т.д.
5. Экспортируйте набор данных в текстовый файл с настройками, предлагаемыми по умолчанию.
6. Импортируйте только что экспортированный файл в Deductor.
7. Присоедините к новому узлу импорта (путем копирования) предыдущую ветвь, начиная с узла Настройка набора данных.
8. Между экспортом и настройкой набора данных вставьте еще один узел настройки, в котором измените тип столбца Поле2 на логический.
9. Сохраните проект.
10. Настройте следующие визуализаторы к любому узлу импорта: Таблица, Статистика. Перейдите в режим формы и обратно. Имеются ли пропуски в записях?
11. В визуализаторе Таблица настройте, чтобы при отображении к значениям в Поле3 добавлялось слово «кг.». Сохраните конфигурацию визуализатора под названием «K1».

12. Сделайте первые три столбца невидимыми. Сохраните конфигурацию визуализатора под названием «K2».

13. Вернитесь к конфигурации K1.

14. В визуализаторе Таблица установите фильтр «Полеб = не пустой». Удалите фильтр.

15. Создайте новый проект. Импортируйте в него текстовый файл CreditSample.txt, идущий в поставке Deductor (по умолчанию расположен в каталоге /Samples директории установки Deductor).

16. Отсортируйте этот набор данных по следующим полям в порядке возрастания: Срок ссуды, Размер ссуды, Количество иждивенцев.

17. Сделайте следующую замену (после Сортировки) в поле Семейное положение: значение Да измените на Женат/замужем, Нет – на Холост/Не замужем.

18. Сделайте следующую замену (после предыдущего узла Замена данных) в поле Количество иждивенцев: значение 0 – на Нет, 1 – без изменений, 2 и 3 – 2 и более. Используйте два способа – непосредственным вводом в мастере обработки и через файл таблицы соответствий. Файл подстановок предварительно создайте в любом текстовом редакторе, например, в Блокноте.

19. Старое поле Количество иждивенцев удалите из набора данных, а новое поле Количество иждивенцев_REPLACE переименуйте в Иждивенцы.

20. Отфильтруйте набор данных, полученный в п. 5 по полю Иждивенцы так, чтобы в выходной набор попали только строки, у которых значение в поле Иждивенцы не равно Нет. Сколько записей прошло через фильтр?

21. Отфильтруйте набор данных, полученный в п. 5 по полю Иждивенцы так, чтобы в выходной набор попали только строки, у которых значение в поле Иждивенцы не равно Н/д. Сколько записей прошло через фильтр?

22. Продолжите фильтровать набор данных, полученный в п. 6. Наложите следующий фильтр, в который попадают все записи, удовлетворяющие условиям а либо условиям б: а. Размер ссуды – от 2000 до 5000, Цель ссуды – Покупка товара; б. Цель ссуды – Иное.

23. Сколько записей прошло через фильтр?

24. Отсортируйте последний набор данных по полю Код.