**Технологии анализа данных.
Методические указания к практическим занятиям**

**Содержание**

[Введение 2](#_Toc410885568)

[Лабораторная работа № 1. Хранилище данных 5](#_Toc410885569)

[Лабораторная работа № 2. ETL 6](#_Toc410885570)

[Лабораторная работа № 3. Срезы OLAP-кубов 8](#_Toc410885571)

# Введение

## Описание предметной области

Компания занимается сборочным производством и продажей устройств и комплектов из деталей, закупаемых у поставщиков. Компания имеет два филиала, которые географически удалены друг от друга и имеют **отличающуюся** информационную структуру.

Аналитик компании выполняет подготовку различных оперативных отчетов, целью которых является увеличение эффективности деятельности компании в целом.

Необходимо разработать программную систему анализа данных для аналитика компании, которая выполняет следующие основные функции: интеграция данных из филиалов в хранилище данных компании, подготовка оперативных отчетов.

Схема баз данных обоих филиалов представлена на Рис. 1.



**Рис. 1.** Схема баз данных филиалов компании.

Описание сущности «Поставщик» представлено в Табл. 1.

**Табл. 1.** Атрибуты сущности Поставщик (S)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Атрибут | Ключ | Семантика | Тип данных |
| 1 | SID | \* | Уникальный код поставщика | INT |
| 2 | SName |  | Имя поставщика | CHAR(20) |
| 3 | SCity |  | Город поставщика | CHAR(20) |
| 4 | Address |  | Почтовый адрес поставщика | CHAR(50) |
| 5 | Risk |  | Риск сотрудничества с поставщиком (низкий, средний, высокий) | INT |

Описание сущности «Поставщик» представлено в Табл. 2.

**Табл. 2.** Атрибуты сущности Деталь (P)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Атрибут | Ключ | Семантика | Тип данных |
| 1 | PID | \* | Уникальный код детали | INT |
| 2 | PName |  | Имя детали | CHAR(20) |
| 3 | PCity |  | Город детали | CHAR(20) |
| 4 | Color |  | Цвет детали | CHAR(20) |
| 5 | Weight |  | Вес детали в килограммах | FLOAT |

Описание сущности «Поставка» представлено в Табл. 3.

**Табл. 3.** Атрибуты сущности Поставка (SP)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Атрибут | Ключ | Семантика | Тип данных |
| 1 | SPID | \* | Уникальный код поставки | INT |
| 2 | SID | ^S.SID | Уникальный код поставщика | INT |
| 3 | PID | ^P.PID | Уникальный код детали | INT |
| 4 | Quantity |  | Количество деталей в поставке | INT |
| 5 | Price |  | Цена за 1 деталь. | FLOAT |
| 6 | ShipDate |  | Фактическая дата доставки | DATE |

Данные в каждом из филиалов должны подчиняться ограничениям целостности, перечисленным в Табл. 4. Тем не менее, в филиалах не всегда осуществляется проверка целостности вводимых данных, вследствие чего в данных **возможны ошибки**.

**Табл. 4.** Ограничения целостности

|  |  |
| --- | --- |
| Сущность | Ограничение целостности |
| S | SName NOT NULL |
| SCity NOT NULL |
| Address NOT NULL |
| (SName, Address, SCity) UNIQUE |
| Risk in (1, 2, 3) |
| P | PName NOT NULL |
| PCity NOT NULL |
| Weight > 0 |
| (PName, PCity, Color) UNIQUE |
| SP | ShipDate NOT NULL |
| Qty > 0 |
| Price >0 |
| SP.Qty \* P.Weight <= 1500 |

В первом филиале для обработки данных используется СУБД MS Access. Во втором филиале для обработки данных используется MS Excel.

Схемы данных в обоих филиалах **соответствуют** схеме данных, представленной на Рис. 1.

## Полезные ссылки

* Курс технологий анализа данных
	+ <http://mzym.susu.ru/courses/dat/index.html>
* Microsoft SQL Server
	+ <https://msdn.microsoft.com/ru-ru/library/bb545450.aspx>
* Microsoft SQL Server Integration Services
	+ <https://msdn.microsoft.com/ru-ru/library/ms141026.aspx>
* Microsoft SQL Server Reporting Services
	+ <https://msdn.microsoft.com/ru-ru/library/ms159106.aspx>

# Лабораторная работа № 1. Хранилище данных

**Количество баллов: 5**.

**ЦЕЛЬ.** Проектирование и создание базы данных в СУБД Microsoft SQL Server, которая будет использоваться в качестве хранилища данных.

1. Скачайте файлы Branch\_1.accdb и Branch\_2.xsls. Откройте их в MS Access и MS Excel соответственно. Убедитесь в верности отображения данных.
2. Спроектируйте схему данных хранилища. При проектировании нужно учесть следующие особенности последующего использования данных:
	* При обработке даты поставки планируется **раздельное** использование дня, месяца и года. Допускается использование отдельной таблицы с датами и внешний ключ в таблице поставок.
	* При обработке веса и стоимости поставки планируется использование только **общего** веса и **полной** стоимости поставки.
	* При обработке поставки планируется использование характеристики веса поставки со значениями из (1, 2, 3), где 1 – легкая поставка, а 3 – тяжёлая поставка.
	* При обработке поставки планируется использование характеристики цены детали в поставке: дешёвая/дорогая.
3. Подключитесь к MS SQL Server, используя следующие параметры:
	* Имя сервера: mssqldb.pvc.susu;
	* Проверка подлинности: проверка подлинности SQL Server;
	* Имя пользователя: student;
	* Пароль: Student1.
4. Создайте базу данных хранилища с именем «ваша\_фамилия\_warehouse» в соответствии с разработанной схемой данных.

# Лабораторная работа № 2. ETL

**Количество баллов: 10**.

**ЦЕЛЬ.** Разработка и реализация алгоритмов ETL для интеграции данных филиалов в хранилище с использованием Microsoft SQL Server Integration Services.

## Задание 1. Реализация ETL для поставщиков и деталей.

**Количество баллов: 4**.

1. Запустите SQL Server Data Tools и создайте новый проект Integration Services.
2. Создайте общий поток управления ETL.
3. Создайте отдельные потоки данных для интегрирования данных поставщиков и для интегрирования данных деталей.
4. Добавьте потоки данных в поток управления и реализуйте их с учетом следующих указаний:
	* После импортирования данных из источников (MS Excel или MS Access) их необходимо очистить (Cleaning) с помощью отброса тех записей, которые не соблюдают ограничения целостности, представленные в Табл. 4.
	**Факультативное задание (1 балл):** при очистке данных исправлять и сохранять записи поставщиков с незаданным полем Risk, устанавливая его в максимальное значение.
	* После слияния двух таблиц возможно невыполнение уникальности ключей (SName, SCity, Addresss) или (PName, PCity, Color). В этом случае необходимо включить в хранилище только первую запись.
	**Факультативное задание (2 балла):** включить в хранилище обе записи, сохранив вторую и последующие с добавлением номера записи в SName или PName соответственно.
5. Создайте скриншоты потока управления, потока данных поставщиков и потока данных деталей для использования в качестве отчёта о выполнении лабораторной работы.

## Задание 2. Реализация ETL для поставок.

**Количество баллов: 6**.

1. Создайте поток данных для интегрирования данных поставок и добавьте его в поток управления, созданный в предыдущем задании данной лабораторной работы.
2. Реализуйте поток данных поставок с учетом следующих указаний и **особенностей спроектированного вами хранилища**:
	* После импортирования данных из источников (MS Excel или MS Access) их необходимо очистить (Cleaning) с помощью отброса тех записей, которые не соблюдают ограничения целостности, представленные в Табл. 4.
	**Факультативное задание (2 балла):** при очистке данных исправлять и сохранять записи поставок с незаданным значением Quantity, выставляя его в среднее значение.
	* Необходимо учесть, что в каждом из филиалов свои идентификаторы поставок, деталей и поставщиков, и сохранить целостность ссылок.
3. Создайте скриншоты потока управления и потока данных поставок для использования в качестве отчёта о выполнении лабораторной работы.

# Лабораторная работа № 3. Срезы OLAP-кубов

**Количество баллов: 5**.

**ЦЕЛЬ.** Создание отчётов с использование Microsoft SQL Server Reporting Services в виде срезов OLAP-кубов.

## Задание 1. Отчёт на основании среза OLAP-куба.

**Количество баллов: 2**.

1. Запустите SQL Server Data Tools и создайте проект Reporting Services.
2. Создайте соединение с источником данных, в качестве которого укажите хранилище данных, созданное и заполненное в рамках предыдущих лабораторных работ.
3. Создайте новый общий набор данных с помощью запроса на основании таблиц поставщиков, поставок и, если была создана отдельная, таблицы Time.
4. Создайте новый пустой отчёт.
5. Используя элемент таблицы создайте отчёт с возможностью консолидации (roll-up) и детализации (drill-down) данных по следующей иерархии:
	* Ось времени: год > месяц > день.
	* Ось поставщиков: город поставщика > название поставщика.

Учитывайте, что для корректного отображения понадобятся дополнительные столбцы и строки с суммарными значениями.
Учитывайте, что вне зависимости от формата хранения времени в базе данных, необходимо корректно отображать названия месяцев.
**Факультативное задание (1 балл):** используйте дополнительные средства оформления отчёта для более удобного и красивого отображения среза.

1. Создайте скриншот полностью детализированного отчёта для использования в качестве отчёта о выполнении задания лабораторной работы.

## Задание 2. Отчёты на основании срезов OLAP-куба.

**Количество баллов: 3**.

1. По алгоритму, использованному в предыдущем задании этой лабораторной работы, создайте отчёт по осям поставщики-детали с возможностью консолидации (roll-up) и детализации (drill-down) данных по следующей иерархии:
	* Ось деталей: город детали > название детали > цвет детали.
	* Ось поставщиков: город поставщика > название поставщика.
2. По алгоритму, использованному в предыдущем задании этой лабораторной работы, создайте отчёт по осям время-детали с возможностью консолидации (roll-up) и детализации (drill-down) данных по следующей иерархии:
	* Ось времени: год > месяц > день.
	* Ось деталей: деталь > вес поставки (легкий, средний, тяжёлый).
3. По алгоритму, использованному в предыдущем задании этой лабораторной работы, создайте отчёт по осям время-детали с возможностью консолидации (roll-up) и детализации (drill-down) данных по следующей иерархии:
	* Ось времени: год > месяц > день.
	* Ось деталей: деталь > цена детали (дорогая, дешёвая).
4. Создайте скриншоты полностью детализированных отчётов для использования в качестве отчёта о выполнении задания лабораторной работы.