МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
"Южно-Уральский государственный университет"
(национальный исследовательский университет)

Образовательная программа по направлению подготовки 010400.62

«Информационные технологии» (степень «бакалавр»)

ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ

ЗАДАНИЕ ПО ВЫПОЛНЕНИЮ КУРСОВОГО ПРОЕКТА

Челябинск-2013

Введение

Целью курсового проекта является разработка приложения интеллектуального анализа данных для модельной предметной области. Документ организован следующим образом.

В первом разделе приведено описание модельной предметной области.

Второй раздел содержит практические задания, результатом последовательного выполнения которых будет разработка приложения для модельной предметной области.

Документ завершается *списком литературы*, рекомендуемой при выполнении заданий.

В *приложение* вынесен шаблон отчета о выполнении курсового проекта.

1. Описание предметной области

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

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

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

Далее в разделе 1.1 приведено описание сущностей предметной области. В разделе 1.2 описаны базы данных филиалов компании. Раздел 1.3 содержит описание функций программной системы, которую необходимо разработать.

1.1. Сущности предметной области

При описании сущностей используются обозначения, указанные в табл. 1.

Табл. 1. Обозначения атрибутов сущностей

Обозначение	**Семантика**
*	Атрибут является первичным ключом сущности
^Сущность.Атрибут	Атрибут является внешним ключом и ссылается на указанный атрибут указанной сущности

В предметной области выделены сущности Поставщик, Деталь и Поставка. Описание сущности Поставщик представлено в табл. 2.

Табл. 2. 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)	(1, 2, 3)

Атрибуты сущности Деталь представлены в табл. 3.

Табл. 3. Атрибуты сущности Деталь (P)

№	Атрибут	Ключ	Семантика	Тип данных
1.	PID	*	Уникальный код детали	INT
2.	PName		Имя детали	CHAR(20)
3.	HTP		Является ли деталь продуктом высоких технологий (High Technology Product)	BOOL
4.	Weight		Вес детали в килограммах	FLOAT

Описание атрибутов сущности Поставка представлено в табл. 4.

Табл. 4. 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.	Qty		Количество деталей в поставке	INT
5.	Price		Цена за 1 шт.	FLOAT
6.	OrderDate		Дата заказа поставки	DATE
7.	Period		Срок доставки в днях	INT
8.	ShipDate		Фактическая дата доставки	DATE

Данные в каждом из филиалов должны подчиняться ограничениям целостности, перечисленным в табл. 5. Тем не менее, в филиалах не всегда осуществляется проверка целостности вводимых данных, вследствие чего в данных возможны ошибки.

Табл. 5. Ограничения целостности

Сущность	Ограничение целостности
S	SName NOT NULL
	SCity NOT NULL
	Address NOT NULL
	(SName, Address, SCity) UNIQUE
P	PName NOT NULL
	Weight > 0
	HTP in (0, 1)
	(PName, Weight) UNIQUE
SP	OrderDate NOT NULL
	Qty > 0
	Price > 0
	Period >= 0
	OrderDate <= ShipDate
S, P, SP	Вес поставки не должен превышать 1,5 тонн

1.2. Описание схем баз данных филиалов

В филиале № 1 для обработки данных используется СУБД MS Access. База данных представляет собой совокупность реляционных таблиц S, P, SP, структура которых описана в разделе 1.1.

В филиале № 2 для обработки данных используются электронные таблицы MS Excel. База данных представляет собой один файл электронных таблиц в формате MS Excel с листами S, P, SP, структура которых описана в разделе 1.1.

В филиалах компании используются унифицированные уникальные идентификаторы поставщиков и деталей, но различные уникальные идентификаторы поставок.

В базе данных как первого, так и второго филиала возможны нарушения ограничений целостности, указанных в табл. 5.

1.3. Описание системы анализа бизнес-данных

Система анализа бизнес-данных должна обеспечивать следующие основные функции:

1. Поддержка хранилища данных.
2. Оперативный анализ данных.
3. Интеллектуальный анализ данных.

Поддержка хранилища данных

Хранилище данных компании используется как основной и единственный источник данных для системы анализа бизнес-данных. Хранилище данных компании создается путем *интеграции* баз данных первого и второго филиалов. Для обозначения процесса интеграции используется термин *ETL (Extract, Transform, Load)*, поскольку данный процесс осуществляется как последовательность следующих шагов: извлечение данных, трансформация и очистка извлеченных данных и загрузка трансформированных и очищенных данных в хранилище.

Очистка подразумевает обработку ошибок в данных. В модельной предметной области ошибочные данные (имеющие нарушения ограничений целостности, указанных в табл. 5), должны быть отброшены. Оставшиеся данные необходимо подвергнуть трансформации для приведения их к форматам данных в хранилище.

Извлечение данных должно осуществляться с периодичностью, задаваемой пользователем системы (например, один раз в неделю в воскресенье). Си-

стема должна обеспечивать также принудительное извлечение данных по запросу пользователя.

В модельной предметной области используются следующие *три измерения*: Время, Место и Товар, каждое из которых имеет два уровня иерархии (см. рис. 1).

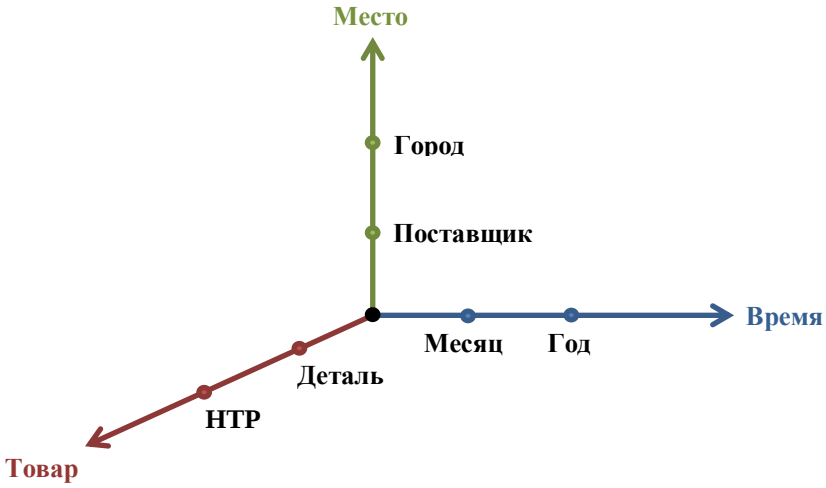


Рис. 1. Измерения предметной области

Измерение *Время* имеет уровни иерархии *Месяц* и *Год*. Измерение *Место* имеет уровни иерархии *Город* и *Поставщик*. Измерение *Товар* имеет уровни иерархии *Деталь* и *НТР* (принадлежность детали к продукту высоких технологий).

В качестве *меры* в модельной предметной области используется сумма поставки, вычисляемая как произведение количества деталей в поставке Qty на цену одной детали Price.

Оперативный анализ данных

Система анализа бизнес-данных должна обеспечивать *оперативный анализ данных (OLAP, Online Analytical Processing)*, который обеспечивает следующие основные функции: визуализация OLAP-куба и (или) его срезов, подготовка различных отчетов с агрегированием информации.

OLAP-куб представляет собой данные хранилища в виде многомерного куба, в котором каждое измерение дополняется специальным значением *ВСЕГО*, и полученные таким образом новые точки пространства вычисляются с помощью заданной *агрегатной функции*.

В модельной предметной области в качестве агрегатной функции используется суммирование. Пример OLAP-куба модельной предметной области представлен на рис. 2.

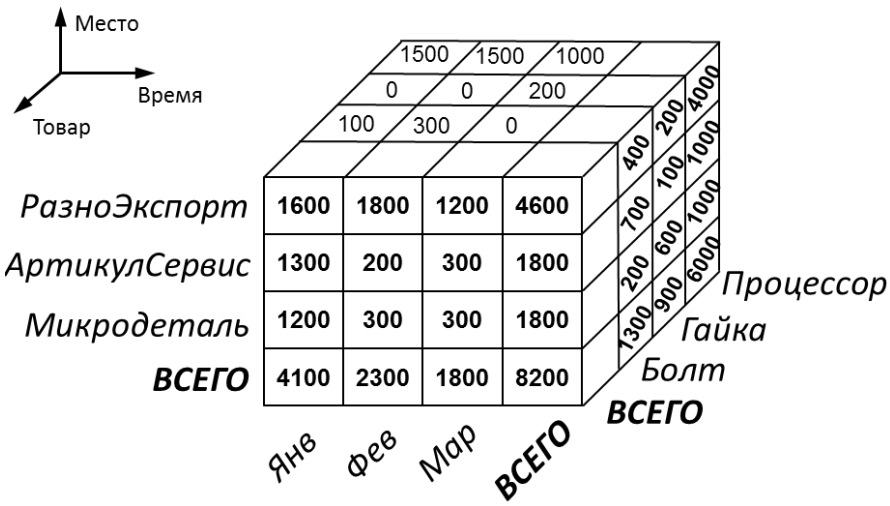
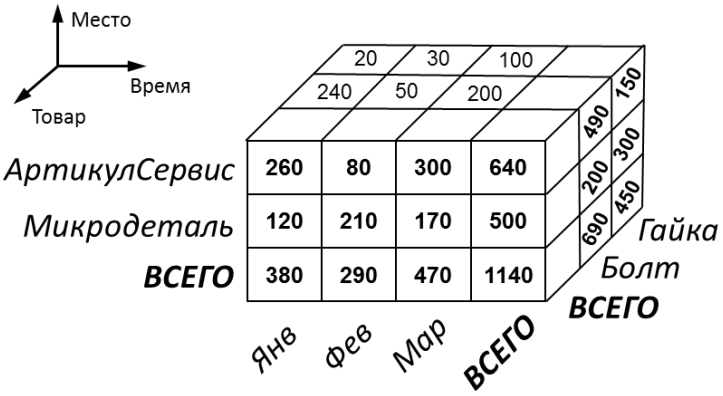


Рис. 2. Пример OLAP-куба

Срез OLAP-куба представляет собой подкуб исходного куба, полученный путем фильтрации данных по одной или нескольким осям. Пример срезов OLAP-куба модельной предметной области представлен на рис. 3.



а) 3-мерный срез OLAP-куба

	Янв	*Фев*	*Мар*	***ВСЕГО***
Процессор	3000	2000	1000	**6000**
Гайка	260	220	320	**900**
Болт	840	80	480	**1300**
ВСЕГО	**4100**	**2300**	**1800**	**8200**

б) 2-мерный срез OLAP-куба

Рис. 3. Пример срезов OLAP-куба

Визуализация OLAP-куба и его срезов предполагает возможность *консолидации (roll-up)* и *детализации (drill-down) данных* в соответствии с уровнями иерархии в измерениях. Пример консолидированных данных приведен на рис. 4. Источником данных этих примеров являются детализированные данные среза OLAP-куба, показанного на рис. 3б.

	Янв	*Фев*	*Мар*	***ВСЕГО***
Высоко-технологичные	3000	2000	1000	**6000**
Низко-технологичные	1100	300	800	**2200**
ВСЕГО	**4100**	**2300**	**1800**	**8200**

Рис. 4. Пример консолидации данных

Интеллектуальный анализ данных

Система анализа бизнес-данных должна обеспечивать *интеллектуальный анализ данных (Data Mining)*, который обеспечивает следующие основные функции: классификация, кластеризация и поиск шаблонов.

Классификация данных предполагает автоматизированное распределение данных по непересекающимся классам, количество и семантика которых заранее известны. Классификация выполняется на основе *обучающей выборки* данных, в которой классы заранее указаны экспертом.

Система анализа бизнес-данных должна обеспечивать интерфейс для выполнения разбиения поставщиков на три категории по риску сотрудничества с ними: низкий, средний, высокий. Для классификации поставщиков должны использоваться следующие атрибуты: общая сумма поставок от данного поставщика, общее количество поставленных данным поставщиком деталей, общее количество фактов срыва поставки данным поставщиком. Срыв поставки имеет место, если фактическая дата доставки позже, чем сумма даты заказа и срока доставки.

Кластеризация данных предполагает автоматизированное распределение данных по непересекающимся кластерам, количество которых заранее извест-

но, а семантика – заранее неизвестна. В отличие от классификации, в кластеризации не участвует эксперт и не используется обучающая выборка.

Система анализа бизнес-данных должна обеспечивать интерфейс для выполнения кластеризации поставок, сделанных в указанный период. Для кластеризации поставок должны использоваться следующие атрибуты: вес поставки, сумма поставки, количество деталей в поставке, название детали, город детали, признак НТР детали в поставке, цена детали в поставке.

Поиск шаблонов направлен на определение наборов данных, которых повторяются с частотой, не ниже указанной. Система анализа бизнес-данных должна обеспечивать интерфейс для поиска наборов деталей, которые часто доставляются совместно в один и тот же день.

2. Практические задания

Данный раздел содержит практические задания, результатом последовательного выполнения которых будет разработка системы анализа бизнес-данных для модельной предметной области.

2.1. Создание источников данных

1. Разработка генератора модельных данных

Разработайте генератор модельных баз данных филиалов компании. Генератор должен представлять собой консольное приложение, которое создает файлы базы данных в формате CSV (Comma-Separated Values – значения, разделенные запятыми) в зависимости от указанных параметров запуска. Пример запуска генератора:

```
gendb -sname s.csv -pname p.csv -spname sp.csv
      -sqty 100 -pqty 100 -spqty 1000
      -serr 10 -perr 10 -sperr 10
```

В параметрах запуска префиксы **s**, **p**, **sp** означают соответствующие таблицы модельной базы данных, префикс **name** – имя файла, префикс **qty** – количество записей в таблице, префикс **err** – процент ошибочных записей (имеющих нарушения ограничений целостности, указанных в табл. 5) в соответствующей таблице.

Названия в генерируемых данных должны быть осмысленными.

2. Создание модельных баз данных

С помощью ранее разработанного генератора создайте базы данных для 1-го и 2-го филиалов компании. Общее количество записей в таблицах не должно быть менее значений (для ошибочных записей – не быть более значений), указанных в примере запуска генератора.

2.2. Разработка пользовательского интерфейса системы

1. Проектирование пользовательского интерфейса

Разработайте главное меню и основные формы пользовательского интерфейса системы анализа бизнес-данных в соответствии с функциями системы, описанными в разделе 1.3.

2. Разработка прототипа системы

Реализуйте прототип системы, заменив реакцию приложения на выбор пользователем еще не разработанных функций выдачей сообщений-"заглушек".

2.3. Разработка хранилища данных

1. Проектирование хранилища данных

Разработайте схему хранилища данных. Каждая таблица хранилища должна иметь полный и не избыточный набор полей с нужными типами и семантикой для реализации задач, описанных в разделе 1.3.

2. Разработка функций ETL

Разработайте подпрограммы, которые будут обеспечивать извлечение, трансформацию, очистку и загрузку данных в хранилище данных.

3. Включение функций ETL в прототип

С помощью разработанных функций ETL замените соответствующие заглушки прототипа реальной функциональностью, после чего создайте хранилище данных системы.

2.4. Разработка функций оперативного анализа данных

1. Выполните реализацию функций визуализации OLAP-куба или/и его срезов.
2. Выполните реализацию функций консолидации и детализации данных.
3. С помощью разработанных функций оперативного анализа данных замените соответствующие заглушки прототипа реальной функциональностью.

2.5. Разработка функций интеллектуального анализа данных

1. Выполните реализацию функций поиска шаблонов. Замените соответствующие заглушки прототипа реальной функциональностью.
2. Выполните реализацию функций кластеризации. Замените соответствующие заглушки прототипа реальной функциональностью.
3. Выполните реализацию функций классификации. Замените соответствующие заглушки прототипа реальной функциональностью.

Литература

1. Барсегян А.А., Куприянов М.С., Холод И.И., Тесс М.Д., Елизаров С.И. Анализ данных и процессов. СПб.: БХВ-Петербург, 2009. 512 с.
2. Han J., Kamber M. Data Mining: Concepts and Techniques, 2nd ed. Morgan Kaufmann Publishers, 2006. 743 p.

Приложение. Шаблон оформления отчета о выполнении курсового проекта

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
"Южно-Уральский государственный университет"
(национальный исследовательский университет)
Факультет Вычислительной математики и информатики
Кафедра системного программирования

КУРСОВОЙ ПРОЕКТ

бакалавра направления 010400.62 "Информационные технологии"
по дисциплине "Технологии анализа данных"

Разработка приложения интеллектуального анализа данных на платформе СУБД <указать использованную СУБД>

Выполнил:
студент группы <группа>
< И.О. Фамилия студента>

Проверил:
< И.О. Фамилия руководителя>
<ученая степень, ученое звание>

Оценка: _____

Подпись: _____

Дата: _____

Челябинск-2013

1. Задание

1.1. Предметная область

В данном пункте следует поместить *краткое* описание предметной области (сущности, схемы баз данных филиалов); при этом допустимо использовать текст разделов 1.1. и 1.2.

1.2. Функции программной системы

В данном пункте следует поместить *краткое* описание основных функций системы анализа бизнес-данных; при этом допустимо использовать текст раздела 1.3.

2. Проектирование[1]

2.1. Схема хранилища данных

В данном пункте следует поместить следующую информацию:

- рисунок или скриншот, на котором указаны таблицы хранилища данных, связи между ними и типы связей;
- описание таблиц хранилища данных (ключи, поля и их семантика) в табличном виде.

2.2. ETL

В данном пункте следует поместить модульную структуру и интерфейсы подпрограмм, реализующих функции извлечения данных из источников, преобразования и очистки данных, загрузки данных в хранилище.

2.2. Функции OLAP

В данном пункте следует поместить модульную структуру и интерфейсы подпрограмм, реализующих функции OLAP.

2.2. Функции интеллектуального анализа данных

В данном пункте следует поместить модульную структуру и интерфейсы подпрограмм, реализующих функции интеллектуального анализа данных.

2.4. Интерфейс пользователя

В данном пункте следует поместить *краткое* описание основных принципов проектирования интерфейса пользователя системы анализа бизнес-данных и дать ссылки на соответствующие рисунки (скриншоты) Приложения.

3. Реализация

3.1. ETL

В данном пункте следует указать язык программирования и инструментальные средства, на основе которых была выполнена реализация функций ETL, а также привести объем созданных лично исходных текстов в строках (с округлением до сотен).

3.2. Функции OLAP

В данном пункте следует указать язык программирования и инструментальные средства, на основе которых была выполнена реализация функций OLAP, а также привести объем созданных лично исходных текстов в строках (с округлением до сотен).

---

[1] Этот и два последующих раздела (в том числе их подразделы) включаются в отчет только в случае выполнения соответствующих работ.

3.3. Функции интеллектуального анализа данных

В данном пункте следует указать язык программирования и инструментальные средства, на основе которых была выполнена реализация функций интеллектуального анализа данных, а также привести объем созданных лично исходных текстов в строках (с округлением до сотен).

4. Тестирование

4.1. Генератор тестовых баз данных

В данном пункте следует кратко описать назначение, интерфейс и модульную структуру генератора тестовых баз данных, а также привести объем созданных лично исходных текстов в строках (с округлением до сотен); при этом допустимо использовать текст раздела 2.1.

4.2. ETL

В данном пункте следует указать параметры запуска генератора тестовых баз данных, для создания хранилища данных, которое использовалось для тестирования функций системы анализа бизнес-данных, и привести количество записей в таблицах хранилища данных.

4.3. Функции OLAP

В данном пункте следует указать семантику запросов, использованных для тестирования функций OLAP, а также дать ссылку на Приложение 2.

4.4. Функции интеллектуального анализа данных

В данном пункте следует указать семантику запросов, использованных для тестирования функций интеллектуального анализа данных, а также дать ссылку на Приложение 3.

5. Заключение

В заключении следует дать перечень основных полученных результатов: какие компоненты системы спроектированы, реализованы, протестированы, каков объем созданных лично исходных текстов в строках (с округлением до сотен).

6. Приложения

Приложение 1. Интерфейс пользователя (основные формы)

В данном пункте следует поместить скриншоты главного меню (главной формы) и основных форм интерфейса пользователя системы анализа бизнес-данных. Скриншот должен быть оформлен как рисунок с номером и подписью.

Приложение 2. Тестирование функций OLAP

В данном пункте следует поместить несколько (два-три) скриншотов отчетов, генерируемых системой анализа бизнес-данных при выполнении функций OLAP. Скриншот должен быть оформлен как рисунок с номером и подписью.

Приложение 3. Тестирование функций интеллектуального анализа данных

В данном пункте следует поместить несколько (два-три) скриншотов очетов, генерируемых системой анализа бизнес-данных при выполнении функций интеллектуального анализа данных. Скриншот должен быть оформлен как рисунок с номером и подписью.