Федеральное государственное бюджетное образовательное учреждение
высшего образования
«КАЛИНИНГРАДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Н. Б. Розен

ETL-СИСТЕМЫ И БАЗЫ ДАННЫХ

Учебно-методическое пособие по изучению дисциплины
основной профессиональной образовательной программы магистратуры по направлению

09.04.01 – ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

Калининград
Издательство ФГБОУ ВО «КГТУ»
2023

Рецензент:
кандидат технических наук, доцент, директор Института цифровых технологий ФГБОУ ВО «Калининградский государственный технический университет» А. Б. Тристанов

Розен, Н. Б.
ETL-системы и базы данных : Учебно-методическое пособие по изучению дисциплины основной профессиональной образовательной программы магистратуры по направлению 09.04.01 – Информатика и вычислительная техника / **Н. Б. Розен.** – Калининград : Изд-во ФГБОУ ВО «КГТУ», 2023. – 30 с.

В учебно-методическом пособии приведены тематический план и методические указания по изучению дисциплины «ETL-системы и базы данных». Рассмотрены рекомендации по подготовке к промежуточной аттестации и критерии оценивания. Пособие подготовлено в соответствии с требованиями утвержденной рабочей программы общепрофессионального модуля по дисциплине «ETL-системы и базы данных» магистратуры направления подготовки 09.04.01 – Информатика и вычислительная техника.
Табл. 1, список лит. – 12 наименований

Учебно-методическое пособие рассмотрено и одобрено в качестве локального электронного методического материала на заседании кафедры прикладной математики и информационных технологий Института цифровых технологий ФГБОУ ВО «Калининградский государственный технический университет» 24 мая 2023 г., протокол № 5.

Учебно-методическое пособие по изучению дисциплины рекомендовано к использованию в учебном процессе в качестве локального электронного методического материала методической комиссией ИЦТ 19 июня 2023 г., протокол № 7.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Дисциплина «ETL-системы и базы данных» предназначена для студентов магистратуры, обучающихся по направлению 09.04.01 – Информатика и вычислительная техника.

Дисциплина Б1.О.04.01 «ETL-системы и базы данных» относится к общепрофессиональному модулю основной профессиональной образовательной программы магистратуры.

Дисциплина изучается в первом семестре первого курса и базируется на знаниях, полученных в рамках высшего профессионального образования.

Для успешного освоения данной дисциплины слушателям требуются знания по дисциплинам:

- «Информационные технологии» в части представлений о техническом и программном обеспечении, организации работы в распределенных системах и сетевых технологиях;
- «Базы данных» в части проектирования и создания баз данных и языка манипулирования данными SQL;
- «Информационные системы» в части знания технологии обработки информации в информационной системе, организации хранилищ данных и их роли в системах поддержки принятия решений.

Знания, умения и навыки, полученные в результате изучения дисциплины, необходимы для успешного освоения следующих дисциплин:

- «Технологии Data Mining» в части знания способов предварительной подготовки данных, современной организации хранилищ данных и использования современных баз данных;
- «Проектирование и разработка систем интеллектуального анализа данных» в части знания навыков загрузки данных в ПО LOGINOM, проведение некоторых видов интеллектуального анализа данных;
- «Проектирование и разработка интеллектуальных систем поддержки принятия решений» в части знания способов трансформации и очистки данных.

В результате изучения дисциплины обучающийся должен знать процесс переноса данных (ETL-процесс), включающий в себя этапы извлечения, преобразования и загрузки данных, а также программные средства баз данных, обеспечивающие его выполнение.

Особенности современной организации хранилища данных, необходимые способы преобразования, очистки данных и способы их трансформации.

Для достижения цели ставятся задачи:

- сформировать представление о месте процесса ETL в архитектуре

системы бизнес-аналитики на основе хранилищ данных;

- изучить основные элементы ETL-процесса;
- сформировать навыки по выполнению общего планирования и проектирования ETL-процесса;
- обеспечить владение инструментами аналитической обработки;
- определить возможности, которые предоставляют современные реляционные СУБД для проектирования хранилищ данных в системах бизнес-аналитики;
- рассмотреть особенности современных баз данных, в том числе PostgreSQL, в работе с хранилищами данных;
- сформировать навыки использования аналитических, статистических и ранжирующих функций в PostgreSQL для работы в хранилищах данных.

Учебно-методическое пособие по изучению дисциплины содержит учебно-тематический план, включающий перечень изучаемых тем, обязательных лабораторных занятий, мероприятий текущей аттестации и список тем, вынесенных на самостоятельную работу. При формировании личного образовательного плана на семестр обучающемуся следует оценивать рекомендуемое время на изучение дисциплины и возможность больших затрат времени на выполнение некоторых заданий или проработку отдельных тем.

В разделе «Содержание дисциплины» приведены подробные сведения о вопросах, рассматриваемых в данном курсе. Представлены методические рекомендации преподавателя для самостоятельной работы. Каждая тема включает ссылку на литературу (или иной информационный ресурс), а также контрольные вопросы для самопроверки и тесты для самодиагностики по изученной теме.

Раздел «Текущая аттестация» содержит описание обязательных мероприятий контроля самостоятельной работы и усвоения разделов или отдельных тем дисциплины. Изложены требования к промежуточной аттестации, проходящей в форме защиты курсовой работы и экзамена.

Помимо данного пособия, студентам следует использовать материалы, размещенные в соответствующем разделе курса по дисциплине «ETL-системы и базы данных» в ЭИОС КГТУ.

1. ТЕМАТИЧЕСКИЙ ПЛАН

	Раздел (модуль) дисциплины	Тема	Объем аудиторной работы, ч	Объем самостоятельной работы, ч
		Лекции		
1.1	**Аналитика больших данных**	Тема 1. Большие данные. Подходы и определения	2	3
1.2		Тема 2. Содержание и задачи процесса управления большими данными	2	3
2.1	**Процесс ETL и его этапы**	Тема 3. Концепция систем складирования данных и хранилищ данных	2	3
2.2		Тема 4. Цель и организация ETL-процесса	2	3
2.3		Тема 5. Метод многомерного моделирования данных для хранилища данных	2	3
3.1	**Особенности разработки хранилищ данных на основе механизмов программного обеспечения СУБД**	Тема 6. Моделирование объектов физической модели хранилища данных	2	3
3.2		Тема 7. Основы проектирования физической структуры хранилища данных на основе СУБД-ориентированных средств	2	3
4.1	**Возможности реляционной базы данных PostgreSQL**	Тема 8. Применение PostgreSQL для работы с хранилищами данных	2	3
			16	**24**
		Лабораторные занятия		
1.1	**Аналитика больших данных**	Знакомство с платформой Loginom	2	3
1.2		Основная терминология платформы Loginom	2	3
1.3		Изучение стандартных компонент системы Loginom	4	3
1.4		Создание сценария системы Loginom	4	3
2.1	**Процесс ETL и его этапы**	Очистка и предобработка данных в системе Loginom	2	2

	Раздел (модуль) дисциплины	Тема	Объем аудиторной работы, ч	Объем самостоятельной работы, ч
2.2		Сортировка, фильтр строк, замена, дата и время	2	2
2.3		Трансформация данных. Группировка и разгруппировка, квантование, скользящее окно	4	2
2.4		Визуализация в АП Loginom	4	2
4.2	**Возможности реляционной базы данных PostgreSQL**	Доступ к данным в базе PostgreSQL	2	2
4.3		Расширенные возможности выборки данных в базе PostgreSQL	4	2
			30	**24**
		Курсовая работа		
1.1	**Аналитика больших данных**	Контрольная точка 1. Выбор темы и изучение источников данных	-	15
2.1	**Процесс ETL и его этапы**	Контрольная точка 2. Построение сценария ETL	-	15
		Оформление проекта. Защита	-	3
			-	**33**
		Рубежный (текущий) и итоговый контроль		
		Итоговый контроль (экзамен)	-	33,75
				33,75
		Всего	**46**	**81**

2. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ И УКАЗАНИЯ К ИЗУЧЕНИЮ

2.1. Раздел 1. Аналитика больших данных

2.1.1. Тема 1.1. Большие данные. Подходы и определения

Перечень изучаемых вопросов:

Сущность понятия «большие данные». Источники больших данных. Управление на основе данных. Значение больших данных для бизнеса и государства. Лучшие практики. Структурированные и неструктурированные данные. Метаданные. Технологии распределенных вычислений. Инструментальные средства работы с большими данными. Методики анализа больших данных.

Методические указания к изучению:

Уделить внимание причинам появления феномена «большие данные». Знать три основные характеристики больших данных, реализующих принцип трех V (Volume, Velocity, Variety). Уметь приводить примеры использования больших данных в различных областях. Понимать и уметь определять, какие данные можно отнести к структурированным и неструктурированным. Познакомиться с информационными платформами, работающими по тематике больших данных. Знать критерии сравнения таких ресурсов. Необходимо знать примеры лучшего опыта реализации проектов в области больших данных в Российской Федерации.

Литература:

[1] раздел 1 п. 1.1–1.2, стр. 6–20; [2] стр. 5–7; [3] стр. 32–35; [4].

Контрольные вопросы:

1. Что означает термин Big Data в информационных технологиях?
2. Что является основной целью обработки Big Data?
3. Каковы главные характеристики Big Data?
4. Какие понятия содержит в себе принцип трех V?
5. Что входит в понятие «структурированные данные»?
6. Какие данные называются неструктурированными?

2.1.2. Тема 1.2. Содержание и задачи процесса управления большими данными

Перечень изучаемых вопросов:

Понятие жизненного пути больших данных и цикла управления ими. Модель «пути» Малькольма Чисхолма и ее семь активных фаз взаимодействия с данными. Специальные инструменты экосистемы больших данных (Hadoop и базы NoSQL), их подход для извлечения, преобразования и загрузки данных.

Проблемы использования больших данных. Алгоритм MapReduce как модель для распределенных вычислений и его стадии. Инструменты MapReduce Hadoop, Spark, Pig, Hive, Cassandra и Kafka.

Методические указания к изучению:

Целесообразно рассматривать развитие основ работы с большими данными в исторической последовательности. Это дает возможность лучше осознать ограничения и возможности каждого этапа и причины трансформации работы организации к организации, «управляемой данными». Большое внимание следует уделить инструментам работы с большими данными, их возможностям и сравнению по разным критериям.

Литература:

[2] стр. 28–32; [7] раздел 1, стр. 5–36.

Контрольные вопросы:

1. Перечислите этапы жизненного пути больших данных и прокомментируйте каждый из них.
2. Опишите особенности инструмента Hadoop?
3. Каковы особенности базы NoSQL?
4. Каковы причины появления алгоритма MapReduce?
5. Опишите стадии обработки данных в MapReduce.
6. Перечислите особенности алгоритма MapReduce.

2.2. Раздел 2. Процесс ETL и его этапы

2.2.1. Тема 2.1. Концепция систем складирования данных и хранилищ данных

Перечень изучаемых вопросов:

Необходимость разделения данных в системах операционной обработки данных и системах анализа данных. Основные предпосылки возникновения и

сферы применения систем складирование данных (data warehousing) и концепция хранилища данных (data warehouse), их сходство и отличие. Понятие базы данных. СУБД. Модели данных. Хранилище данных, витрины (киоски) данных. Зависимость структуры хранилища данных от структуры бизнеса. Основы языка SQL. Денормализация модели данных.

Методические указания к изучению:

Подчеркнуть место концепции хранилищ данных как ключевого компонента информационной инфраструктуры и архитектуры аналитических приложений для различных сфер в производстве, науке и технологиях. Представить существующий рынок систем бизнес-аналитики. Подчеркнуть роль бизнес-аналитики и, в частности, проектирования хранилищ данных.

Некоторые вопросы темы ранее изучались студентами в дисциплинах «Информационные технологии» и «Информационные системы», но требуют уточнения и изменения акцентов изложения. Например, при рассмотрении вопросов использования баз данных большое внимание уделяется построению реляционной модели данных. Вместе с тем полностью нормализованная модель может быть очень неэффективной при реализации в хранилище данных. Поэтому при преобразовании нормализованной логической модели в физическую модель допускают значительную денормализацию.

Литература:

[2] стр. 28–32; [3] стр. 16–18; [4]; [5]; [6]; [7] раздел 2.

Контрольные вопросы:

1. Назовите отличия хранилища данных от базы данных.
2. Дайте определение понятия «интегрированность» с точки зрения организации хранилища данных.
3. Дайте определение понятия «привязка ко времени» с точки зрения организации хранилища данных.
4. Дайте определение понятия «неизменяемость» с точки зрения организации хранилища данных.
5. Определите понятие «денормализация».

2.2.2. Тема 2.2. Цель и организация ETL-процесса

Перечень изучаемых вопросов:

Возможности переноса данных внешних источников в хранилище систем бизнес-аналитики. Организация процесса ETL как составная часть проекта

разработки хранилища данных. Этапы базового процесса ETL: «Извлечение данных», «Очистка данных», «Трансформация», «Загрузка».

Свойства процесса ETL: большой объем данных, выбираемый из систем источников данных, периодичность процесса ETL, формирование метаданных хранилища. Требование обеспечения качества и контроля данных, поступающих в хранилище, а также восстанавливаемости после сбоев без потери данных.

Методические указания к изучению:

Следует обратить внимание на разные подходы к реализации ETL-процесса: с использованием промежуточной области, без использования промежуточной области, преобразованию данных с использованием сервера хранилища данных, в процессе их загрузки, и таким образом обозначить место процедур разработки и планирования ETL-процесса.

Литература:

[2] стр. 19–20; [3]; [4]; [5]; [6]; [7].

Контрольные вопросы:

1. Определите цель ETL-процесса.
2. Сформулируйте свойства ETL-процесса.
3. Чем определяется периодичность ETL-процесса.
4. Что является метаданными хранилища?
5. Какие подходы к реализации ETL-процесса вам известны?

2.2.3. Тема 2.3. Метод многомерного моделирования данных для хранилища данных

Перечень изучаемых вопросов:

Основные элементы метода многомерного моделирования: факты, атрибуты, измерения, параметры (метрики), иерархия. Понятие пространства признаков. Метрики. Введение в сокращение размерности. Сокращение числа признаков. Оценка качества данных. Технологии и методы оценки качества данных. Очистка и предобработка. Фильтрация данных. Обработка дубликатов и противоречий. Выявление аномальных значений. Восстановление пропущенных значений.

Визуальное представление многомерной модели с помощью гиперкуба. Операции «Развертка» (drill down) и «Свертка» (drill up).

Методические указания к изучению:

Важно понимать, что соблюдение многочисленных правил позволит достичь требуемого качества информации. Обнаружение и устранение ошибок и несоответствий в данных должно быть разным при работе с одним источником или при интеграции нескольких источников данных. Важно обеспечить следование столбцов в одном порядке и проверить, чтобы данные находились в одном формате (например, дата и валюта). Произвести обогащение данных путем объединения дополнительной информации, если это необходимо. Очистку данных следует выполнять не изолированно, а вместе с преобразованием данных, связанных со схемой, на основе всеобъемлющих метаданных. Желательно пояснять все правила соответствующими примерами.

Литература:
[2] стр. 25–36; [7].

Контрольные вопросы:

1. Определите сходство и отличие многомерного моделирования и моделирования «сущность-связь».
2. Поясните смысл понятий «факты» и «атрибуты» с точки зрения многомерного моделирования.
3. Поясните смысл понятия «параметры» с точки зрения многомерного моделирования.
4. Поясните смысл понятия «иерархия» с точки зрения многомерного моделирования.
5. Поясните смысл понятия «гранулированность» с точки зрения многомерного моделирования.

2.3. Раздел 3. Особенности разработки хранилищ данных на основе механизмов программного обеспечения СУБД

2.3.1. Тема 3.1. Моделирование объектов физической модели хранилища данных

Перечень изучаемых вопросов:

Особенности реляционных СУБД: построчное хранение данных, индексирование записей и журнализация транзакций. Колоночные СУБД как повышение эффективности работы аналитических систем. Нормализация схемы реляционной БД и ее влияние на производительность запросов приложений и хранимых объемов информации. Многомерная схема моделирования систем хранилищ данных разработанных для аналитических целей (OLAP). Основные

типы многомерных схем: схема «Звезда», схема «Снежинка» и схема «Галактика».

Методические указания к изучению:

Технологии, которые используют реляционные СУБД и хранилища данных, были разработаны для решения разных задач. Основная задача баз данных – надежное хранение больших объемов информации для учетных систем и при необходимости – быстрый поиск в них. Это обусловило архитектурные особенности реляционных СУБД (построчное хранение данных, индексирование записей и журнализацию операций с помощью специальных файлов). Технологически это выполняется за счет значительного увеличения объема базы данных на диске.

Аналитические информационные системы и хранилища данных, применяющиеся для управленческого анализа и накопленных в учетных системах сведений, работают по другому принципу. Необходимо небольшое количество выборок больших объемов записей, часто с выполнением группировок и расчетом итоговых значений, но с небольшим количеством полей. Это делает использование стандартной СУБД неэффективной и требует поиска совершенно других решений, которые частично реализуются колонковыми СУБД.

Литература:

[2] стр. 21–25; [7]; [8]; [9].

Контрольные вопросы:

1. Каким образом происходит журнализация в базах данных?
2. Объясните особенности организации колоночных СУБД, их достоинства и недостатки.
3. Что такое нормализация базы данных?
4. Определите плюсы и минусы многомерной схемы «Звезда».
5. Определите плюсы и минусы многомерной схемы «Снежинка».
6. Определите плюсы и минусы многомерной схемы «Галактика».

2.3.2. Тема 3.2. Основы проектирования физической структуры хранилища данных на основе СУБД-ориентированных средств

Перечень изучаемых вопросов:

Повышение производительности запросов с помощью построения индексов. Индекс со структурой B-Tree. Свойства индексов со структурой B-Tree. Кластеризованный индекс. Хэш-индексы. Факторы, влияющие на

эффективность индексов. Понятие о секционировании таблиц. Повышение производительности запросов на основе секционирования. Понятие о кластеризации. Возможности SQL для работы с индексами.

Методические указания к изучению:

Можно рекомендовать разделить изучение данной темы на несколько этапов. Первоначально желательно разобраться с назначением индексов. Затем следует изучить причины появления различных способов построения индексов, а также познакомиться с положительными и отрицательными сторонами каждого типа индексов. Стоит указать, что неправильный выбор принципа индексирования может привести к значительному снижению производительности запросов к хранилищу. Наконец, следует понимать, что в каждом типе программного обеспечения средства работы с индексами весьма различны и стоит представлять себе их возможности.

Литература:

[3]; [4]; [5]; [6]; [7].

Контрольные вопросы:

1. Определите понятие индекса.
2. Сформулируйте, что собой представляет индекс физически.
3. Определите назначение и способ построения индекса со структурой B-Tree.
4. Определите назначение и способ построения кластеризованного индекса.
5. Назначение и особенности хэш-индексов.

2.4. Раздел 4. Возможности реляционной базы данных PostgreSQL

2.4.1. Тема 4.1. Применение PostgreSQL для работы с хранилищами данных

Перечень изучаемых вопросов:

Архитектура PostgreSQL. Понятие транзакции. Типы данных. Обзор основных операторов. Возможности индексирования. Обзор применения различных типов индексов: Hash, B-tree, GiST, SP-GiST, GIN и RUM, BRIN и Bloom.

Методические указания к изучению:

Необходимо сформировать представление о возможностях современного инструмента по работе с хранилищами данных. Представлять простые способы работы для выполнения стандартных операций. Понимать, что применение PostgreSQL полезно там, где требуется работа со сложными структурами данных, с которыми не справляются простые СУБД.

Литература:

[11]; [12]; [13].

Контрольные вопросы:

1. Определите назначение и возможности PostgreSQL.
2. Охарактеризуйте понятие «транзакция» с точки зрения системы PostgreSQL?
3. Перечислите типы данных, применяемые в PostgreSQL.
4. Возможности индексирования в PostgreSQL.

3. ТРЕБОВАНИЯ К АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ

Формирование компетенций обеспечивается проведением лекционных занятий и лабораторных работ, в ходе которых приобретаются умения и навыки. Этому же способствуют ответы на контрольные вопросы, самостоятельная работа, ее контроль, выполнение и защита курсовой работы.

Проверка знаний осуществляется с помощью выполнения тестовых заданий. Типовые тесты, предназначенные для самопроверки, приведены в Приложении 1.

Аттестация по дисциплине предполагает постоянный контроль преподавателем качества усвоения учебного материала, активизацию учебной деятельности на занятиях, побуждение их к самостоятельной систематической работе. Их результаты учитываются выставлением оценок в ходе ежемесячной аттестации.

Практически на всех занятиях применяется выборочный контроль, который позволяет понять, в какой степени усвоен материал.

Обязательным условием допуска к экзамену является выполнение и защита лабораторных работ и защита курсовой работы, предусмотренные учебным планом по данной дисциплине.

Примерный перечень тем курсовых работ приведен в Приложении 6.

Вопросы для подготовки к экзаменам приведены в Приложении 2, а образцы билетов к экзаменам – в Приложении 3.

Подготовка к экзамену ведется на основе конспекта лекций, рекомендуемым к изучению учебникам и учебным пособиям. В ходе подготовки к экзамену преподаватель проводит консультацию, на которой доводится порядок проведения экзамена и даются ответы на вопросы, вызвавшие наибольшие затруднения в процессе подготовки.

Экзамен является заключительным этапом изучения дисциплины. Он проводится в объеме рабочей программы. Для проведения экзамена разработаны экзаменационные билеты. В экзаменационный билет включены два теоретических вопроса из разных разделов программы и одно практическое задание. Слушатели заранее знакомятся с вопросами к экзамену. Предварительное ознакомление слушателей с экзаменационными билетами не разрешается.

Для подготовки к ответу слушателям отводится порядка 30 минут. По окончании ответа экзаменатор может задавать дополнительные и уточняющие вопросы в пределах учебного материала, вынесенного на экзамен. Прерывать экзаменующегося во время ответа не рекомендуется.

Оценка по результатам экзамена объявляется слушателю, заносится в экзаменационную ведомость и зачетную книжку. Неудовлетворительные оценки проставляются только в экзаменационной ведомости (в зачетные книжки не заносятся). Неявка на экзамен отмечается в экзаменационной ведомости: «не явился». Другие записи или прочерки в экзаменационной ведомости не допускаются.

Пересдача экзамена допускается не более одного раза. При получении повторной неудовлетворительной оценки окончательное решение об уровне подготовленности слушателя принимает специальная комиссия. Слушатели, знания которых оценены комиссией как неудовлетворительные, отчисляется из вуза.

Знания, умения и навыки слушателей при текущем и промежуточном контроле определяются оценками: «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Критерии оценки знаний:

«Отлично» – слушатель глубоко и прочно усвоил весь учебный материал, исчерпывающе, последовательно, грамотно и логически стройно его изложил, может ответить при видоизменении задания, свободно справляется с задачами и практическими заданиями, правильно обосновывает принятые решения, умеет самостоятельно обобщать и излагать материал, не допуская ошибок.

«Хорошо» – слушатель знает программный материал, грамотно и по существу излагает его, не допускает существенных ошибок в ответе на вопросы, может правильно применять теоретические положения и владеет необходимыми

умениями и навыками при выполнении практических заданий.

«Удовлетворительно» – слушатель усвоил основной материал, но не глубоко, допускает неточности, нарушает последовательность в изложении программного материала и испытывает затруднения в выполнении практических заданий.

«Неудовлетворительно» – слушатель не знает значительной части программного материала, допускает существенные ошибки, с большими затруднениями выполняет практические задания, задачи.

Оценка («отлично», «хорошо», «удовлетворительно» или «неудовлетворительно») выставляется в соответствии с критериями, указанными в таблице 1:

Таблица 1

Критерий	Система оценок			
	2	3	4	5
	0–40 %	41–60 %	61–80 %	81–100 %
	неудовлетворительно	удовлетворительно	хорошо	отлично
Системность и полнота знаний в отношении изучаемых объектов	Обладает частичными и разрозненными знаниями, которые не может научно корректно связывать между собой (только некоторые из них может связывать между собой)	Обладает минимальным набором знаний, необходимым для системного взгляда на изучаемый объект	Обладает набором знаний, достаточным для системного взгляда на изучаемый объект	Обладает полнотой знаний и системным взглядом на изучаемый объект

4. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО САМОСТОЯТЕЛЬНОЙ РАБОТЕ

4.1. Цели и задачи самостоятельной работы

К современному специалисту предъявляются жесткие требования, среди которых – возможность самостоятельно приобретать знания, выбирать их из различных источников, систематизировать полученную информацию, давать ей оценку. Формирование навыков происходит в течение всего периода обучения: при выполнении практических занятий и тестов, написании курсовой работы и особенно при самостоятельной работе.

Целью самостоятельной работы по специальности 09.04.01 – Информатика и вычислительная техника при изучении дисциплины «ETL-системы и базы данных» является овладение знаниями в области подготовки данных, профессиональными умениями и навыками применения этих знаний в будущей специальности, опытом творческой, исследовательской деятельности.

Самостоятельная работа способствует развитию ответственности и организованности, творческого подхода к решению проблем учебного и профессионального уровня.

Задачами самостоятельной работы студентов являются: систематизация и закрепление полученных теоретических знаний и практических умений; углубление и расширение теоретических знаний; развитие навыков использования нормативной, правовой, справочной документации и специальной литературы; формирование самостоятельности мышления; развития навыков проведения самостоятельного исследования.

4.2. Рекомендации по организации самостоятельной работы студентов магистратуры с лекционным материалом

Лекции являются важной формой учебного процесса, так как способствуют получению знаний и освоению новых методов изучения материала. Они позволяют упростить восприятие нового материала, устанавливать связь учебного материала со специальностью, знакомят с новейшими научными достижениями в области информации и информационных технологий.

Дисциплина «ETL-системы и базы данных» читается в первом семестре первого курса магистратуры. С первых занятий студент получает большое количество теоретической информации и практических заданий. Для успешного усвоения знаний и выполнения заданий необходима четкая организация самостоятельной работы, прежде всего правильное планирование своего времени. За основу для планирования рекомендуется тематический план и вопросы по самостоятельной работе, представленные в данном документе.

Большое количество времени, выделенное на самостоятельную работу, предполагает возможность планомерного и целенаправленного изучения лекционного материала, его многократное повторение, что обеспечивает надежное закрепление в памяти.

Лекции необходимо изучать систематически, в течение всего семестра. При первом чтении изучается весь материал, рассматриваются базовые положения, заучиваются определения и формулы. При втором обычно достаточно рассмотреть только отдельные важные положения, а в дальнейшем повторяются лишь отдельные определения. Такая работа упрощается при наличии качественного конспекта лекций.

Конспект лекций не является единственным источником информации при изучении дисциплины. В программе дисциплины «ETL-системы и базы данных» указан подробный список рекомендуемых источников для изучения.

Если возникли вопросы при рассмотрении лекционного материала или при изучении дополнительных источников, то рекомендуется обсудить их в аудитории в специально отведенное для ответов на вопросы время.

Некоторые теоретические разделы дисциплины специально выносятся для самостоятельного изучения во время внеаудиторной работы. В этом случае необходимо дополнить конспект. Обычно преподаватель дает ссылки на учебники, пособия или другие источники в которых можно почерпнуть эти сведения.

Примерный перечень вопросов для самостоятельной работы с теоретическим материалом приведен в Приложении 4.

4.3. Самопроверка

После изучения определенной темы по записям в конспекте и учебнику, а также лабораторных работ рекомендуется воспроизвести по памяти определения и формулировки основных положений.

В случае недостаточного уровня усвоения необходимо вернуться к ранее пройденному материалу.

Полезно пройти тестирование. Примеры тестовых заданий приведены в Приложении 1.

Следует избегать механического заучивания формулировок и попыток выполнения лабораторных работ без понимания сущности применяемой технологии.

Примерные вопросы для самопроверки приведены в Приложении 5.

4.4. Рекомендации по организации самостоятельной работы при подготовке курсовых работ

Выполнение курсовой работы относится к одному из видов самостоятельной работы. Курсовая работа способствует:

- углублению и расширению знаний;
- формированию интереса к познавательной деятельности;
- овладению приемами процесса познания;
- формированию представления о месте дисциплины «ETL-системы и базы данных» в будущей специальности.

Курсовая работа является исследовательской работой. При ее подготовке следует пользоваться стандартным планом выполнения исследовательской работы. Рекомендуется начать с изучения особенностей современного состояния проблемы и подобрать необходимые теоретические источники.

4.5. Рекомендации по организации самостоятельной работы при подготовке к экзаменам

Экзамен – форма заключительной проверки знаний, умений, навыков, степени развития обучающихся в системе образования.

Главная задача состоит в том, чтобы у студента в результате подготовки к экзамену из отдельных сведений и деталей сформировалось представление об общем содержании дисциплины, стала понятной философия предмета, его система. Готовясь к экзамену, слушатель приводит в систему знания, полученные на лекциях и лабораторных занятиях, разбирается в том, что осталось непонятным. Это позволит на экзамене свободно ответить на теоретические вопросы, показать умение применять полученные знания на практике, связать их с использованием в будущей специальности.

На экзамене оцениваются:

- понимание и степень усвоения теории;
- знание рекомендуемой литературы, современных публикаций;
- умение использовать теорию на практике, решать конкретные задачи;
- логика, структура и стиль ответа,
- умение пояснять выдвигаемые положения.

Примерный перечень вопросов на экзамене приведен в Приложении 2.

5. ЗАКЛЮЧЕНИЕ

В учебно-методическом пособии нашли отражение следующие характеристики дисциплины:

1) тематический план;

2) содержание дисциплины и указания к ее изучению, включающие по каждой теме: перечень изучаемых вопросов, методические указания к изучению темы, ссылки на литературу, контрольные вопросы;

3) рекомендации по самостоятельной работе;

4) требования к аттестации по дисциплине: содержание текущей аттестации, условия получения положительной оценки на экзамене, примерные вопросы к экзамену.

6. БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Прокопенко, Н. Ю. Аналитические информационные системы поддержки принятия решений [Текст] : учеб. пособие / Н. Ю. Прокопенко; Нижегор. гос. архитектур.-строит. ун-т. – Н. Новгород : ННГАСУ, 2020. – 142 с.
2. Вольфсон, М. Б. Анализ данных : учебное пособие / М. Б. Вольфсон. – Санкт-Петербург : СПбГУТ им. М. А. Бонч-Бруевича, 2015. – 81 с. –

Текст : электронный // Лань : электронно-библиотечная система. – URL: https://e.lanbook.com/book/180254 (дата обращения: 18.06.2023).
3. Ремарчук, В. Н. Информационная аналитика: теория, методология, технологии / В. Н. Ремарчук. – 2-е изд., стер. – Санкт-Петербург : Лань, 2023. — ISBN 978-5-507-45840-0. –Текст : электронный // Лань : электронно-библиотечная система. – URL: https://e.lanbook.com//book/288980 (дата обращения: 18.06.2023).
4. Официальный сайт компании Loginom Company [Электронный ресурс]. – Режим доступа: https://loginom.ru/ (дата обращения: 04.06.2023).
5. Энциклопедия по бизнес-анализу [Электронный ресурс]. – URL: https://wiki.loginom.ru (дата обращения: 04.06.2023).
6. Loginom: Руководство пользователя [Электронный ресурс]. – Режим доступа: https://help.loginom.ru/userguide/ (дата обращения: 04.06.2023).
7. Талипов, Н. Г. Технологии интеллектуального анализа данных : учебное пособие / Н. Г. Талипов. – Казань : КНИТУ-КАИ, 2020. – 308 с. – ISBN 978-5-7579-2488-5. – Текст : электронный // Лань : электронно-библиотечная система. – URL: https://e.lanbook.com/book/193530 (дата обращения: 24.06.2023). – Режим доступа: для авториз. пользователей.
8. Нестеров С. А. Интеллектуальный анализ данных с использованием SQL Server / С. А. Нестеров. – Санкт-Петербург : Лань, 2023. – ISBN 978-5-507-45535-5. – Текст : электронный // Лань: электронно-библиотечная система. – URL: https://e.lanbook.com/book/311861 (дата обращения: 30.05.2023). – Режим доступа: для авториз. пользователей. – С. 47).
9. Кузнецова, С. В. Информационное обеспечение, базы данных: лабораторные работы : учебное пособие / С. В. Кузнецова. – Москва : МАИ, 2022. – ISBN 978-5-4316-0978-7. – Текст : электронный // Лань : электронно-библиотечная система. – URL: https://e.lanbook.com/book/298634 (дата обращения: 30.05.2023). – Режим доступа: для авториз. пользователей. – С. 2).
10. Сенченко, П. В. Организация баз данных : методические указания / П. В. Сенченко. – Москва : ТУСУР, 2022. – 80 с. – Текст : электронный // Лань : электронно-библиотечная система. – URL: https://e.lanbook.com/book/313088 (дата обращения: 30.05.2023). — Режим доступа: для авториз. пользователей.
11. Наместников, А. М. Базы данных. Практический курс : учебное пособие : в 2-х частях / А. М. Наместников. – Ульяновск : УлГТУ, 2017. – Часть

1 : Объектно-реляционные базы данных на примере PostgreSQL 9.5. – 2017. – 113 с. – ISBN 978-5-9795-1743-8. – Текст : электронный // Лань : электронно-библиотечная система. – URL: https://e.lanbook.com/book/165100 (дата обращения: 24.06.2023). – Режим доступа: для авториз. пользователей.

12. Джуба, С. Изучаем PostgreSQL 10 / С. Джуба, А. Волков. – Москва : ДМК Пресс, 2019. – 400 с. – ISBN 978-5-97060-643-8. – Текст : электронный // Лань : электронно-библиотечная система. – URL: https://e.lanbook.com/book/116125 (дата обращения: 24.06.2023). – Режим доступа: для авториз. пользователей.

ПРИЛОЖЕНИЕ 1

Тестовые задания по дисциплине

Вопрос 1. Закончите последовательность единиц измерения объемов информации: байт, килобайт, мегабайт, гигабайт, терабайт, _____, _____,_____,_____

Вопрос 2. Расшифруйте аббревиатуру ETL
_____, _____,_____,

Вопрос 3. Дополните определение «Хранилище данных» это
_____

Вопрос 4. Существует два основных метода обогащения данных — _____ и _____

Вопрос 5. Продолжите определение: Транзакция это –
_____
_____

Вопрос 6. Перечислите четыре основные характеристики *Big Data*:
а) Virtualization, Volume, Variability, Vehicle
б) Variety, Velocity, Volume, Value
в) Verification, Volume, Velocity, Visualization
г) Video, Value, Variety, Volume

Вопрос 7. Автоматизировать запуск пакетных задач в рамках конвейера обработки больших данных по расписанию можно с помощью…
a. Apache Hadoop
b. Apache Kafka
c. Apache AirFlow
d. Apache Hive

Вопрос 8. Для полнотекстового интеллектуального поиска и аналитики по полуструктурированным данным в формате JSON подходит СУБД…
a. Elasticsearch
b. Hive
c. Cassandra
d. HBase

Вопрос 9. Технология потоковой обработки событий в режиме реального времени …

a. Spark Streaming
b. MapReduce
c. Apache Hadoop
d. Apache Kafka

Вопрос 10. Для машинного обучения подходят данные….

a. Бинарные
b. Предварительно подготовленные, очищенные от ошибок, пропусков и выбросов, а также нормализованные и представленные в виде числовых векторов
c. Любых форматов в цифровом виде мегабайт
d. Числовые типа int

Вопрос 11. Формат Parquet считается…

a. колоночным (столбцовым)
b. полуструктурированным
c. неструктурированным
d. строковым

Вопрос 12. Метод, который разбивает данные на множество небольших наборов тестовых данных, которые могут быть использованы для модификации модели, …

a. Перекрестная проверка
b. Регуляризация
c. Ранняя остановка
d. Переоснащение

Вопрос 13. В базе данных PostgreSQL в результате выполнения оператора:

INSERT INTO software (name, version, release_date, price, company, language, type, platform, description)

VALUES

('Windows 10', '2023', '2019-11-12', '199.99', 'Microsoft', 'English', 'Operating System', 'Windows', 'The latest version of the Microsoft Windows operating system.')

в добавляемой записи поле «version» примет значение _____ .

Вопрос 14. Таблица измерений должна находиться в отношении «_____» с таблицей фактов

Вопрос 15. При использовании схемы «Звезда» центральной является таблица _____, с которой связаны все таблицы _____.

ПРИЛОЖЕНИЕ 2

Примерные вопросы к экзамену

1. Понятие «Большие данные» (Big Data). Роль цифровой информации в XX веке.
2. Характеристики Big Data.
3. Проблемы анализа и обработки Big Data.
4. Основные принципы обработки Big Data.
5. Шкалы измерений. Характеристики и иерархия.
6. Метаданные
7. Технологии обработки данных
8. Понятие структурированных и неструктурированных данных.
9. Технологии обработки больших данных: NoSQL.
10. Технологии обработки больших данных: MapReduce.
11. Технологии обработки больших данных: Hadoop, R.
12. Общая характеристика процесса ETL и его этапов.
13. Системы складирования данных и хранилище данных. Сходство и отличие.
14. Очистка данных, общие сведения и этапы.
15. Предобработка данных. Виды предобработки.
16. Машинное обучение.
17. Понятие гиперкуба. Операции «Развертка» и «Свертка».
18. Многомерная схема моделирования хранилищ данных.
19. Типы многомерных схем: схема «Звезда», схема «Снежинка» и схема «Галактика».
20. Индексы, назначение, особенности построения.
21. Понятие транзакции.
22. Возможности и назначение системы LOGINOM.
23. Применение PostgreSQL.

ПРИЛОЖЕНИЕ 3

Образец шаблона билета для экзамена

Дисциплина:	ETL-системы и базы данных	Специальность:	09.04.01
Семестр:	1 семестр		
Кафедра:	ПМИТ		
1.	Общая характеристика процесса ETL и его этапов		
2.	Представление многомерной модели с помощью гиперкуба		
3.			

ПРИЛОЖЕНИЕ 4

Перечень вопросов для самостоятельной работы

№ п/п	№ раздела дисциплины	Тематика самостоятельной работы *(детализация)*	Контроль выполнения работы
1.	**Раздел 1** Аналитика больших данных	1. Информация и особенности ее хранения и обработки. 2. История развития больших данных. 3. Инструменты управления большими данными	Опрос
2	**Раздел 2** Процесс ETL и его этапы	1. Метрики классификации. 2. Метрика регрессии. 3. Настройка алгоритмов для оптимизации моделей. 4. Оценка моделей, основанная на их точности	Опрос, тест
3	**Раздел 3** Особенности разработки хранилищ данных на основе механизмов программного обеспечения СУБД	1. Разные подходы к архитектуре хранилищ данных. 2. Архитектура корпоративной системы хранилища – *DWH*. 3. Хранилище данных и озеро данных	Опрос, тест
4	**Раздел 4** Возможности реляционной базы данных PostgreSQL	1. Возможности PostgreSQL по работе с геоданными. 2. Виды нереляционных БД	Опрос, тест

ПРИЛОЖЕНИЕ 5

Вопросы для самопроверки

1. Приведите определение термина «большие данные». Перечислите признаки, характеризующие большие данные.
2. Перечислите возможные источники больших данных. Приведите примеры генерации больших данных.
3. Применение больших данных в отраслях. Приведите примеры применения больших данных в областях энергетики, горнодобывающей и нефтяной промышленности, здравоохранении, логистике и транспорте.
4. Приведите примеры лучшего опыта реализации проектов в области больших данных в зарубежных странах.
5. Приведите примеры лучшего опыта реализации проектов в области больших данных в Российской Федерации.
6. Перечислите основные проблемы/сложности в хранении больших данных.
7. Приведите определение термина «машинное обучение». Приведите и дайте определение типам машинного обучения.
8. Общая постановка задачи обучения по прецедентам в теории «Машинного обучения».
9. Типология задач обучения по прецедентам в теории «Машинного обучения».
10. Опишите последовательность операций технологического процесса определения, подготовки и анализа данных.
11. Опишите сетевую часть этапа подготовки данных для анализа.
12. Что представляет собой этап моделирования данных?
13. Опишите цикл моделирования данных и его этапы.
14. Что представляет собой этап проектирования признаков в процессе моделирования?
15. В чем состоит метод перекрестной проверки данных?
16. Каким образом происходит настройка гиперпараметров?
17. Дайте определение точности и прецизионности в прогнозировании.
18. В каких случаях возникает необходимость использования специальных архитектур для обработки больших данных в облачных вычислениях? Приведите примеры.
19. Преимущества и недостатки применения специализированных архитектур для обработки больших данных в облачных вычислениях.

\

ПРИЛОЖЕНИЕ 6

Примерная тематика курсовых работ

1. Разработка ETL-системы на примере данных анализа текста. Очистка данных на примере.
2. Предварительный анализ данных и построение признаков в задачах визуализации информации.
3. Предварительный анализ данных и построение признаков в задачах распознавания темы текста.
4. Прогнозирование и анализ выручки предприятия.
5. Технологии обработки больших данных для извлечения информации о событиях из проблемно-ориентированных текстов.
6. Технологии обработки больших данных для обработки терминологической информации из научно-технических текстов.
7. Технологии обработки больших данных для анализа респондентов социологического исследования, проведенного в Интернете.
8. Разработка системы сбора данных с платформ по поиску вакансий.
9. Технологии обработки больших данных для составления предложений по аренде жилья.
10. Проектирование ETL-системы для сбора информации о местах вузов.
11. Технологии обработки данных для агрегации данных международной статистики из различных источников.
12. Построение и наполнение базы данных статистической информации на основе открытых источников.
13. Построение Excel-приложения для работы с разнородной статистической информацией.
14. Построение инструмента интерактивной визуализации статистической информации.
15. Обработка больших наборов данных о биржевой активности.
16. Проверка статистических гипотез на больших наборах данных о биржевой активности.
17. Классификация больших корпусов текстовых документов.
18. Поиск заимствований в больших корпусах текстовой информации.
19. Разработка системы сбора структурированных данных с интернет-форумов.
20. Сбор и структурирование данных новостного потока.
21. Разработка системы сбора и структурирования данных о номенклатуре определенной группы товаров интернет-магазина.

Локальный электронный методический материал

Нина Борисовна Розен

ETL-СИСТЕМЫ И БАЗЫ ДАННЫХ

Редактор М. А. Дмитриева

Уч.-изд. л. 1,4. Печ. л. 1,9.

Издательство федерального государственного бюджетного
образовательного учреждения высшего образования
«Калининградский государственный технический университет».
236022, Калининград, Советский проспект, 1.