Содержание

Введение

Big Data – это очень важная в современном мире it-отрасль, которая так или иначе задействована практически во всех сферах разработки, поэтому понимание основ работы с базами данных очень важно для любого программиста.

В ходе производственной практики необходимо ознакомиться с основными понятиями, связанными с Big Data, основами их проектирования и управления ими. Одной из главных задач также является изучение современных подходов и методов, используемых при работе с базами данных, такие как ETL разработка.

Место прохождения практики – АО «Неофлекс Консалтинг» – компания, фокусирующаяся на заказной разработке программного обеспечения и внедрении сложных информационных систем. В компании используются передовые технологии и подходы, с некоторыми из которых студентов ознакомят за время практики.

1 Постановка задачи

Главная цель производственный практики заключается в том, чтобы за отведенное время получить базовые знания, необходимые для работы с системой управления базами данных (СУБД) PostgreSQL, написания основных SQL-запросов, проектирования и реализации ETL (Extract, Transform, Load) процессов, проектирования и создания баз данных, а также получить опыт работы в команде.

Определены следующие задачи:

1) Знакомство с языком PL/SQL и его основными возможностями;
2) Изучение понятия СУБД и знакомство с PostgreSQL;
3) Основы проектирования реляционных баз данных;
4) Изучение понятия Data Warehouse (задачи и цели, принципы организации);
5) Изучение ETL-процессов (основные понятия, проектирование и разработка);
6) Проверка собственной базы данных из дипломной работы на соответствие стандартам качества современной разработки.

2 Реализация практических задач

2.1 Изучение языка PL/SQL

Был прослушан курс теоретического материала, посвященного данному языку.

Structured Query Language (SQL) – язык структурированных запросов, с помощью которого к базе данных создаются специальные запросы (или SQL инструкции) с целью получения данных и/или манипулирования ими.

Фактически язык SQL представляет собой набор операторов, посредством которых осуществляются любые действия над данными. Эти операторы принято делить на четыре группы по своему назначению:

1) Data Definition Language (DDL) – отвечает за определение данных. С помощью данной группы операторов происходят все манипуляции над схемой базы данных. Например, создание таблиц, изменение типов данных, удаление таблиц;

2) Data Manipulation Language (DML) – предназначена для манипулирования данными. Данная группа отвечает за получение информации из базы данных, её редактирование и удаление, а также за добавление новых данных;

3) Data Control Language (DCL) – предназначена для определения прав доступа к данным;

4) Transaction Control Language (TCL) – предназначена для управления транзакциями. Транзакция представляет собой некоторый блок команд, которые объединены в одну логическую единицу работы с данными.

Очень важное значение для базы данных имеет понятие Create, Read, Update, Delete (CRUD). Данный термин обозначает четыре базовые операции, используемые при работе с базами данных.

Для добавления новых данных в PL/SQL используется оператор INSERT, для получения (чтения) данных – оператор SELECT, для обновления данных –

оператор UPDATE, а для удаления – оператор DELETE. Фактически данный набор операторов представляет группу DML.

CRUD постоянно используется для всего, что связано с базами данных их проектированием. Разработчики ничего не смогут сделать без CRUD, поэтому при изучении языка PL/SQL в рамках производственной практики особое внимание было уделено изучению и практическому применению этих операций.

2.2 Понятие СУБД и система PostgreSQL

Для дальнейшей работы с базами данных необходимо ознакомиться с понятием СУБД и изучить PostgreSQL.

Реляционная база данных – это некоторый набор данных, представленный в виде таблиц, между которыми предопределены некоторые связи. Связь служит для отображения зависимостей одних наборов данных (таблиц) от других.

Объектно-реляционная база данных – это база данных, которая реализует некоторые технологии объектно-ориентированного программирования (классы, наследование, полиморфизм).

Система, предназначенная для создания и управления объектно-реляционными базами данных называется объектно-реляционная система управления базами данных (ОРСУБД). Одной из таких систем является PostgreSQL.

К основным преимуществам PostgreSQL относят:

1) Поддержку баз данных (БД) неограниченного размера;
2) Расширяемую систему встроенных языков программирования;
3) Мощные и надежные механизмы работы с данными;
4) Легкую расширяемость.

В основном компании выбирают PostgreSQL за большой функционал, возможность эффективно работать с огромным количеством данных и пригодность системы для реализации сложных SQL-запросов.

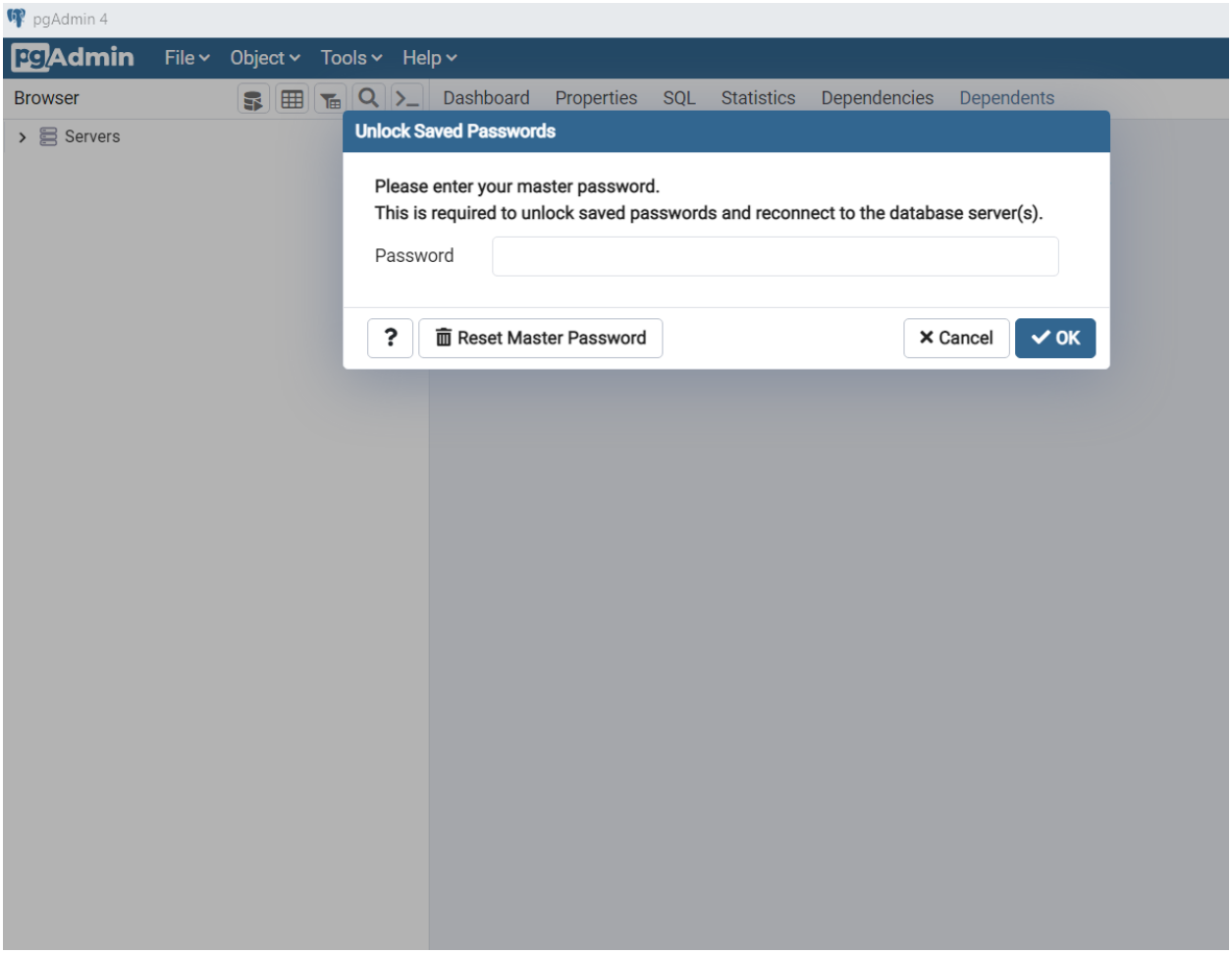

Рис. 2.2.1. Интерфейс pgAdmin, программы для управления PostgreSQL

В теоретическом курсе были рассмотрены основные возможности данной системы, такие как создание новых подключений к базам данных, открытие терминала для написания SQL-команд, интерфейс программы и его особенности.

2.3 Основы проектирования баз данных

Данная тема очень важна для любого, кто работает с базами данных, поскольку понимание основных принципов их проектирования способствует понимаю всех процессов, которые происходят с ними в результате деятельности разработчиков.

Проектированием базы данных называют процесс представления некоторой информации в виде сущностный (таблиц) и определение между ними отношений согласно требованиям поставленной задачи. Основными задачами разработчика при проектировании баз данных являются:

1) Обеспечение хранения всей необходимой информации;

2) Обеспечение возможности получения всей необходимой информации из базы данных;

3) Обеспечение целостности базы данных. Целостностью базы данных называют соответствие имеющейся в неё информации её внутренней структуре и логике;

4) Сокращение избыточности данных.

Выделают три основных этапа проектирования базы данных:

1) Концептуальное проектирование. Предполагает построение модели самого высокого уровня абстракции. На этом этапе разработчик определяет логику построения базы данных, требуемые сущности и связи между ними;

2) Логическое проектирование. Разработчик строит конкретную схему базы данных, определяя для каждой сущности конкретные атрибуты и их типы;

3) Физическое проектирование. Разработчик строит по полученной схеме SQL-скрипт, призванный обеспечить полный перенос логической модели в конкретную СУБД.

При изучении данной темы особое внимание было уделено конкретным примерам проектирования простейших реляционных баз данных, поскольку невозможно качественно изучить данную тему без конкретных практических задач.

Также отдельным понятием, связанным с проектированием баз данных, называют секционирование данных. Под секционированием в общем случае понимают подход к проектированию БД. Основная идея данного подхода –

разделение таблицы на несколько частей меньшего размера. Различают два подвида — горизонтальное и вертикальное секционирование.

Вертикальное секционирование подразумевает вынос отдельных атрибутов (столбцов) базы данных в отдельную таблицу.

При горизонтальном секционировании в отдельные таблицы выносятся строки.

Данный поход позволяет эффективно структурировать данные, а также обеспечить быстрый доступ к требуемой информации.

2.4 Основы Data Warehouse

Data Warehouse (DWH) является очень значимым, неотъемлемым элементом устройства любой крупной it-компании. Данный термин используется для обозначения единого корпоративного хранилища данных из разных источников. Его главной целью является обеспечение компании возможностью принимать верные решения в ключе управления бизнесом посредством предоставления целостной информационной картины.

DWH отличается от обычной базы данных по следующим параметрам:

1) Обычные базы данных хранят в себе информацию о конкретных информационных системах компании. В DWH же хранится информация, получаемая ото всех систем компании;

2) Обычные базы данных сосредоточены на хранении только актуальной информации, в то время как DWH используется как некоторый архив, куда стекаются все исторические данные;

3) DWH играет другую роль в бизнес-процессах. Изначально новая информация поступает в БД, и только оттуда в DWH.

Выделяют четыре главных принципа проектирования DWH:

1) Проблемно-предметная ориентация. Данные объединяются в категории и хранятся в соответствии с областями, которые они описывают, а не с приложениями, которые они используют;

2) Интегрированность. Данные объединены так, чтобы они удовлетворяли всем требованиям предприятия в целом, а не единственной функции бизнеса;

3) Некорректируемость. Данные в хранилище данных не создаются: то есть поступают из внешних источников, не корректируются и не удаляются;

4) Зависимость от времени. Данные в хранилище точны и корректны только в том случае, когда они привязаны к некоторому промежутку или моменту времени.

Поскольку DWH содержит в себе огромное количество самых различных данных, одной из главных задач при работе с хранилищем является их грамотное секционирование.

2.5 ETL процессы

Конечной целью производственной практики являлась реализация собственного простейшего ETL процесса, поэтому данной теме было уделено особенное внимание. ETL (Extract, Transform, Load) – один из ключевых процессов в управлении хранилищами данных, который включает в себя:

1) Извлечение данных из внешних источников;

2) Их очистка и модификация в соответствии с требования бизнес-логики;

3) Последующая загрузка данных в хранилище.

Процессы ETL очень важны для любой системы, работающей с базами данных, поскольку именно посредством их информация переносится в хранилище как из внешних источников, так и из самого хранилища.

Обычно ETL-процесс реализуется в сторонней среде разработки, поскольку язык PL/SQL не лучшим образом показывает себя при изменении

структуры файлов и данных, что необходимо при любом ETL-процессе. В рамках производственной практики реализация ETL-процесса велась на языке Python.

Разработка ETL-процесса включает в себя следующие этапы:

1) Планирование ETL-процесса, которое включает в себя разработку диаграммы потоков данных от систем-источников, определение преобразований, метода генерации ключей и последовательности операций для каждой таблицы назначения;

2) Конструирование процесса заполнения таблиц измерений, которое включает в себя разработку и верификацию процесса заполнение статических таблиц измерений, разработку и верификацию механизмов изменения для каждой таблицы измерений;

3) Конструирование процесса заполнения таблиц фактов, которое включает в себя разработку и верификацию процесса первоначального заполнения и периодического дополнения таблиц фактов, построение агрегатов и разработку процедур автоматизации процесса ETL.

2.6 Проверка собственной базы данных

Под руководством куратора была произведена проверка собственной базы данных, используемой в дипломной работе.

Был сделан вывод, что база данных полностью соответствует современным стандартам разработки и качественно реализует всю бизнес-логику приложения. В дополнительных доработках база данных не нуждается.

Заключение

В ходе производственной практики были приобретены и актуализированы знания по теме «Базы данных», что поспособствовало расширению кругозора и лучшему понимаю данной it-сферы. Обучаемые были ознакомлены с современными подходами к работе с базами данных, а также получили опыт решения практических задач.

Все поставленные цели были выполнены в полном объеме.

Список использованных источников

1) Что такое DDL, DML, DCL и TCL в языке SQL. – URL: https://info-comp.ru/what-is-ddl-dml-dcl-tcl (дата обращения: 21.05.2022)

2) PostgreSQL : Документация. – URL: https://postgrespro.ru/docs/postgresql (дата обращения: 21.05.2022)

3) Проектирование баз данных. – URL: https://ru.wikipedia.org/wiki/Проектирование_баз_данных (дата обращения: 20.05.2022)

4) Что такое Data Warehouse (DWH) и зачем крупному бизнесу корпоративное хранилище данных. – URL: https://cloud.mts.ru/cloud-thinking/blog/data-warehouse (дата обращения: 20.05.2022)