Завдання лабораторних робіт

Лабораторна робота №0: Підготовча робота

Це “формальна” лабораторна робота, що вимагає виконати важливі базові дії щодо налаштування середовища для роботи, виконати встановлення необхідних інструментів розробки та відновити у пам’яті знання з реляційних баз даних, SQL, програмування мовою Python. Для виконання подальших ЛР може виникати необхідність встановлювати додаткові елементи середовища розробки, проте фокус у цій ЛР на тих, які мають бути вже добре відомими.
Мета: Відновити у пам’яті знання щодо роботи з базами даних, написання програмного коду. Нагадати навички розгортання середовища для розробки, включно із підключенням системи контролю версій.
Постановка задачі: Встановити усі необхідні елементи в середовищі розробки (IDE, RDBMS, version control) та перевірити себе написавши нескладний програмний код, що задіює усі елементи створеного середовища.
Завдання: Встановити СУБД PostgreSQL. Встановити середовище для програмування мовою Python (PyCharm від JetBrains, або налаштувати роботу з Python у MS Visual Studio). Підключити систему контролю версій (на вибір будь-яку, GitHub, BitBucket, GitLab тощо). Створити просту БД та наповнити її згенерованими даними. Написати програмний код мовою Python для виконання запитів на вибірку, зміну, видалення даних у створеній БД.

Лабораторна робота №1: Зберігання та доступ до даних для аналітики

Ця лабораторна робота дає легкий старт з простого перетворення файлів, щоб ознайомитися з базовими операціями вводу-виводу та маніпулюванням даними. Практичний результат: перетворення CSV на Parquet та спостереження за впливом на розмір файлу та продуктивність простих запитів.
Мета: Зрозуміти практичний вплив різних форматів файлів на ефективність зберігання даних та продуктивність простих запитів. Дослідити алгоритми стиснення та стовпчастого та рядкового форматів зберігання на високому рівні.
Постановка задачі: Переробити великий CSV-файл з даними (наприклад, фінансових транзакцій) в оптимізований формат для аналітичних запитів.
Завдання: Перетворити дані CSV в Parquet, застосувати стиснення та виконати прості запити для порівняння розміру файлу та продуктивності (використати мову Python).

Лабораторна робота №2: Моделювання даних та проектування схем

Спираючись на попередню лабораторну роботу, присвячену структурованим даним, ця лабораторна робота представляє більш абстрактне та концептуальне завдання проектування схем. Ідея ЛР у тому, щоб перейти від простого перетворення даних до міркувань про те, як логічно структурувати дані для конкретних випадків використання, щоб усунути розрив між необробленими даними та складними запитами.

Мета: Зрозуміти компроміси між різними парадигмами моделювання даних та те, як проектування схем впливає на продуктивність запитів. Ознайомлення з нормалізацією та денормалізацією, а також логічною структурою схем «зірка» та «сніжинка».
Постановка задачі: Розробити та впровадити нову схему для команди бізнес-аналітики, якій потрібно виконувати швидкі, спеціальні запити до даних про продажі.
Завдання: Створити схему «зірка» та схему «сніжинка». Завантажити дані в обидві та виконати аналітичні запити для порівняння продуктивності.

Лабораторна робота №3: Побудова відмовостійкого конвеєра даних

Ця лабораторна робота ознайомлює зі складністю реальних конвеєрів даних, додаючи елемент відмови. Необхідно поєднати свої навички програмування з логічним мисленням, щоб вирішувати такі проблеми, як пошкоджені та дубліковані дані. Це значний крок уперед порівняно з попередніми лабораторними роботами, оскільки вимагає врахування поведінки системи з часом та під навантаженням.
Мета: Розробити та реалізувати конвеєр даних, стійкий до збоїв, забезпечуючи цілісність даних шляхом перевірки та надійної обробки помилок. Ознайомлення з ідемпотентністю, обробкою помилок та теорією надійності в практичному контексті.
Постановка задачі: Побудувати конвеєр для отримання та перевірки ненадійних потокових даних датчиків.
Завдання: Створити генератор даних, який вводить помилки, а потім створити компоненти конвеєра для отримання, перевірки та обробки помилок.

Лабораторна робота №4: Розподілена обробка даних та паралелізм

Це заключна лабораторна робота являє собою найскладніше завдання, що поєднує всі попередні концепції: великомасштабні дані, оптимізацію продуктивності та відмовостійке проектування. Необхідно застосувати свої знання до справді великої проблеми даних, використовуючи розподілений фреймворк. Основна увага приділяється не базовому програмуванню, а розумінню нюансів поведінки розподіленої системи.

Мета: Оволодіти концепціями розподілених обчислень, зосереджуючись на тому, як розподіл даних та операції перетасування впливають на продуктивність та масштабованість паралельних алгоритмів. Проаналізувати продуктивність паралельного алгоритму за допомогою моделей, таких як закон Амдала, та обміркуйте, як обробляти “перекоси” даних (data skew).
Постановка задачі: Проаналізуйте величезний набір даних лог-файлу (занадто великий для однієї машини), щоб обчислити середню тривалість сеансу.
Завдання: Використайте Apache Spark для реалізації аналізу. Порівняйте продуктивність різних стратегій розподілу, щоб спостерігати за впливом перетасування.