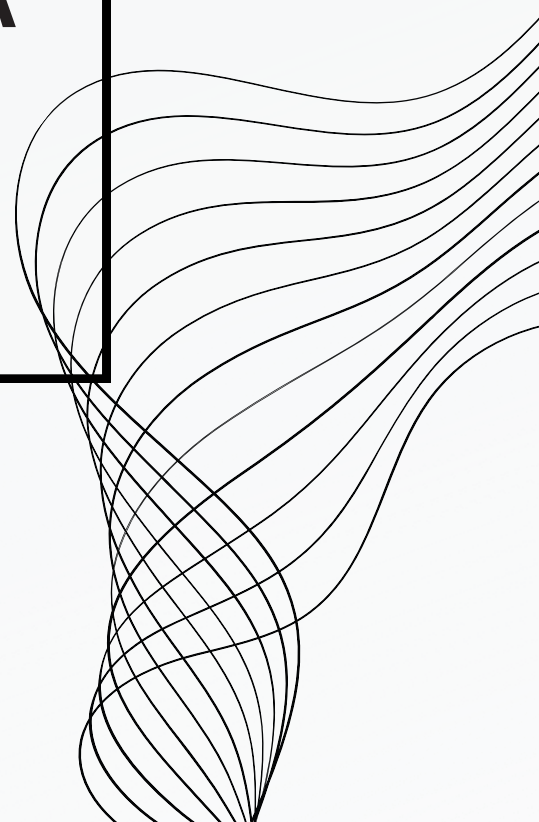




ВВЕДЕННЯ ДО ТЕОРІЇ ХМАРНИХ ОБЧИСЛЕНЬ

**ЛЕКЦІЯ 2. РОЗРОБЛЕННЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ
ДЛЯ АНАЛІЗУ ВЕБ-САЙТІВ, ЯКІ НАДАЮТЬ ВІДКРИТІ ДАНІ ЗА
ДОПОМОГОЮ PYTHON PANDAS. ВІДКРИТІ ДАНІ, ЇХ
ФОРМАТИ ТА ЗАСОБИ ОБРОБЛЕННЯ**



AGENDA

01

МОЖЛИВОСТІ ІНСТРУМЕНТІВ АНАЛІЗУ ДАНИХ

02

РОЛЬ PYTHON В АНАЛІЗІ ДАНИХ

03

ТРАДИЦІЙНА АНАЛІТИКА ВЕЛИКИХ ДАНИХ ТА АНАЛІТИКА
НОВОГО ПОКОЛІННЯ

04

ЖИТТЄВИЙ ЦИКЛ АНАЛІЗУ ДАНИХ

05

ВІДКРИТІ ДАНІ, ЇХ ФОРМАТИ ТА ЗАСОБИ ОБРОБКИ

06

ВЕБ-СКРЕПІНГ

07

ВИТЯГУВАННЯ, ПЕРЕТВОРЕННЯ ТА ЗАВАНТАЖЕННЯ ДАНИХ

2.1. МОЖЛИВОСТІ ІНСТРУМЕНТІВ АНАЛІЗУ ДАНИХ



Вибір інструментів для аналізу залежить від потреби: Для успішного аналізу даних важливо обирати інструменти, які відповідають конкретним вимогам завдання, чи то маніпуляція даними, чи математичне моделювання для прогнозування.



Важливість простоти використання: Легкість вивчення та використання інструменту є ключовим фактором для ефективності. Простота сприяє зниженню часу на навчання та підтримку.



Необхідність маніпуляції даними та спільного доступу: Інструменти для аналізу даних повинні дозволяти маніпулювати даними, очищати їх та забезпечувати можливість спільного доступу до однакових даних для співпраці та однакового їх інтерпретування. Також важливо використовувати інтерактивну візуалізацію для кращого розуміння динаміки даних.

2.2. РОЛЬ PYTHON В АНАЛІЗІ ДАНИХ



Python як ключовий інструмент для оброблення даних: Мова програмування Python стала надзвичайно популярною у сфері обробки даних завдяки своїй легкості вивчення, багатим бібліотекам та великій гнучкості, що дозволяє програмістам ефективно маніпулювати даними та виконувати аналіз.



Використання відкритих інструментів для обробки даних: Багато компаній та організацій обирають використовувати відкриті інструменти для форматування, очищення, аналізу та візуалізації даних, що сприяє зменшенню складності та збереженню часу, який раніше витрачався на навчання та інтеграцію різних програм.



основні бібліотеки Python для аналізу даних.

- NumPy – ця бібліотека додає підтримку роботи з масивами та матрицями. Має багато вбудованих математичних функцій для використання в наборах даних.

- Pandas – ця бібліотека додає підтримку таблиць та часових рядів.

Використовується для маніпулювання та очищення даних.

- Matplotlib – ця бібліотека додає підтримку візуалізації даних. Це бібліотека для простих та складних 3D та контурних графіків.

2.3. ТРАДИЦІЙНА АНАЛІТИКА ВЕЛИКИХ ДАНИХ ТА АНАЛІТИКА НОВОГО ПОКОЛІННЯ



Зростання ролі часу в аналітиці даних: В епоху Big Data час стає критичним фактором для аналізу даних, оскільки дані генеруються в реальному часі, і бізнес повинен бути готовим вживати негайні заходи на основі цих даних.



Переваги рішень, керованих даними: Аналітика даних, орієнтована на час, може покращити рентабельність інвестицій для бізнесу шляхом збільшення часу на дослідження, підвищення ефективності та швидкості виготовлення та входу на ринок, а також поліпшення маркетингу та реклами.



Виклики обробки та аналізу великих обсягів даних: Зі збільшенням обсягу, швидкості та різноманітності даних стає складніше зберігати, обробляти та агрегувати дані за допомогою традиційних аналітичних інструментів.



Значення передбачуваного аналізу в аналітиці наступного покоління: Для прийняття оптимальних рішень у сучасному світі потрібно використовувати інструменти передбачуваного аналізу, які здатні аналізувати велику кількість даних та покращувати точність прогнозування.

АНАЛІТИКА НОВОГО ПОКОЛІННЯ



2.4. ЖИТТЄВИЙ ЦИКЛ АНАЛІЗУ ДАНИХ



Методології аналізу даних: Існує різноманітність методологій для проведення аналізу даних, з популярним стандартом CRISP-DM та індивідуально розробленими методами, які використовуються відповідно до потреб аналітиків.



Життєвий цикл аналізу даних: Життєвий цикл аналізу даних розроблений для використання в бізнес-середовищі та підкреслює можливість ітерацій та необхідність впевненості перед переходом до наступних кроків.



Розподіл методів серед аналітиків: Розподіл використання різних методологій серед аналітиків даних свідчить про широкий спектр підходів та індивідуальний підхід до аналізу даних у відповідь на конкретні завдання.



Значення ітерацій у процесі аналізу даних: Життєвий цикл аналізу даних підкреслює важливість повторних ітерацій та перевірки результатів перед прийняттям остаточних рішень.



ЖИТТЄВИЙ ЦИКЛ АНАЛІЗУ ДАНИХ



2.5. ВІДКРИТІ ДАНІ, ЇХ ФОРМАТИ ТА ЗАСОБИ ОБРОБКИ



Внутрішні дані організацій і джерела їх створення: Для організацій внутрішні дані створюються за допомогою різноманітних систем, включаючи CRM, системи управління навчанням, системи HR та інші.



Необхідність універсального формату файлу: Різні програми створюють файли в різних форматах, що вимагає наявності універсального формату файлу для забезпечення сумісності та обміну даними.



CSV як стандартний формат файлу: Файли CSV (comma-separated values) представляють собою текстовий формат з використанням коми замість крапки для розділення стовпців та символу нового рядка для розділення рядків, і це один із широко використовуваних форматів файлів.



Сумісність JSON і XML: JSON і XML, як інші текстові формати, забезпечують стандартний спосіб представлення даних, що дозволяє їхнє використання в різних сценаріях і сприяє об'єднанню даних з різних джерел.

2.6. ВЕБ-СКРЕПІНГ



Призначення веб-сторінок і процес веб-скрепінгу: Веб-сторінки призначені для людей, але процес веб-скрепінгу використовується для автоматичного видобування даних з HTML-сторінок, схожого на роботу пошукового павука.



Автоматизований процес веб-сканування і збір даних: Веб-сканування включає в себе використання ботів або веб-сканерів для збору та копіювання даних з Інтернету до бази даних або електронної таблиці, щоб їх можна було аналізувати.



Приклади успішного використання відкритих даних: Сайти, як Портал відкритих даних Нью-Йорка та Види веб-скрепіння і їх застосування: Веб-скрепіння може використовуватися для видобування контактів, списків нерухомості, погодних даних, досліджень та порівняння цін. Великі постачальники веб-сервісів надають API для автоматичного збору даних. Garminder, демонструють, як відкриті дані можуть бути використані для забезпечення доступу до інформації та сприяти розумінню соціальних та екологічних питань.



Використання API RESTful для веб-скрепіння: Один із найпоширеніших підходів – використання API RESTful, які використовують HTTP для зв'язку та JSON для кодування даних. Це дозволяє отримувати доступ до великої кількості даних, які генеруються на Інтернет-сайтах, таких як Google та Twitter.

2.7. ВИТЯГУВАННЯ, ПЕРЕТВОРЕННЯ ТА ЗАВАНТАЖЕННЯ ДАНИХ



ETL процес в базах даних: Етапи видобування, перетворення та завантаження (ETL) є ключовими у процесі обробки даних в базах даних. Цей процес передбачає вилучення даних з різних джерел, їх очищення та підготовку для зберігання у бази даних.



Різноманітність джерел та форматів даних: Дані можуть надходити з різних джерел, таких як документи Word, електронні таблиці, текстові файли, презентації, електронні листи та PDF-файли. Ці дані можуть мати різні формати та зберігатися на різних серверах.



Підготовка даних для аналізу: Підготовка даних до аналізу зазвичай займає від 50% до 90% часу, необхідного для проведення аналізу даних. Цей етап включає в себе очищення, трансформацію та консолідацію даних з різних джерел.



Важливість структури та формату даних: Типи та формати даних мають велике значення, коли працюємо з даними в мовах програмування, таких як Python або R. Крім того, різні мови можуть використовувати різні символи для представлення одних і тих самих даних.

2.7. ВИТЯГУВАННЯ, ПЕРЕТВОРЕННЯ ТА ЗАВАНТАЖЕННЯ ДАНИХ



Специфіка структур даних: Структури даних, такі як сховища даних, бази даних, таблиці та рядки, грають важливу роль у зберіганні та організації даних для подальшого аналізу. Програмування Python використовує різні структури даних, такі як списки, словники та дата-фрейми для роботи з даними.



Процеси ETL для об'єднання даних: Оскільки дані можуть зберігатися в різних базах даних, важливим етапом є об'єднання даних з різних джерел у єдиний набір даних для подальшого аналізу.



Важливість стандартизації даних: Стандартизація даних, зокрема формату часу та дати, є важливою для аналізу даних, особливо для роботи з часовими рядами та спостереженнями за динамікою даних.



Значення ETL у процесі аналізу даних: ETL є критичним етапом у підготовці даних для аналізу та використовується для забезпечення якості та зрозумілості даних перед їх використанням у бізнес-аналітиці та наукових дослідженнях.

2.7. ВИТЯГУВАННЯ, ПЕРЕТВОРЕННЯ ТА ЗАВАНТАЖЕННЯ ДАНИХ



Специфіка структур даних: Структури даних, такі як сховища даних, бази даних, таблиці та рядки, грають важливу роль у зберіганні та організації даних для подальшого аналізу. Програмування Python використовує різні структури даних, такі як списки, словники та дата-фрейми для роботи з даними.



Процеси ETL для об'єднання даних: Оскільки дані можуть зберігатися в різних базах даних, важливим етапом є об'єднання даних з різних джерел у єдиний набір даних для подальшого аналізу.



Важливість стандартизації даних: Стандартизація даних, зокрема формату часу та дати, є важливою для аналізу даних, особливо для роботи з часовими рядами та спостереженнями за динамікою даних.



Значення ETL у процесі аналізу даних: ETL є критичним етапом у підготовці даних для аналізу та використовується для забезпечення якості та зрозумілості даних перед їх використанням у бізнес-аналітиці та наукових дослідженнях.

ВИСНОВОК ДО ЛЕКЦІЇ 2

Дані сьогодні не доцільно зберігати на кількох машинах та обробляти лише одним інструментом. Спеціалісти, що приймають рішення, все частіше покладаються на аналітику даних, щоб витягнути необхідну інформацію в потрібний час, у потрібному місці та прийняти правильне рішення.

Прогнозна аналітика передбачає результати та пропонує курси дій, які матимуть найбільшу користь для організації. Файли, дані з мережі Інтернет, датчики та бази даних – це приклади джерел даних. Витягування, перетворення та завантаження (ETL) – це процес збору даних з різних джерел, перетворення даних, а потім завантаження даних у базу даних.

