ВВЕДЕННЯ ДО ТЕОРІЇ ХМАРНИХ ОБЧИСЛЕНЬ

ЛЕКЦІЯ 1. ДЖЕРЕЛА ВЕЛИКИХ ДАНИХ. ІНТЕРНЕТ РЕЧЕЙ. ВИЗНАЧЕННЯ BIG DATA

AGENDA

01 ІНТЕРНЕТ РЕЧЕЙ ТА ЗРОСТАННЯ ДАНИХ

02 ПЛАТФОРМА KAGGLE

03 DRIVENDATA

04 ВИЗНАЧЕННЯ ВЕЛИКИХ ДАНИХ

05 ПРИКЛАДИ ВЕЛИКИХ ДАНИХ У РЕАЛЬНОМУ СВІТІ

06 ВІДКРИТІ ДАНІ

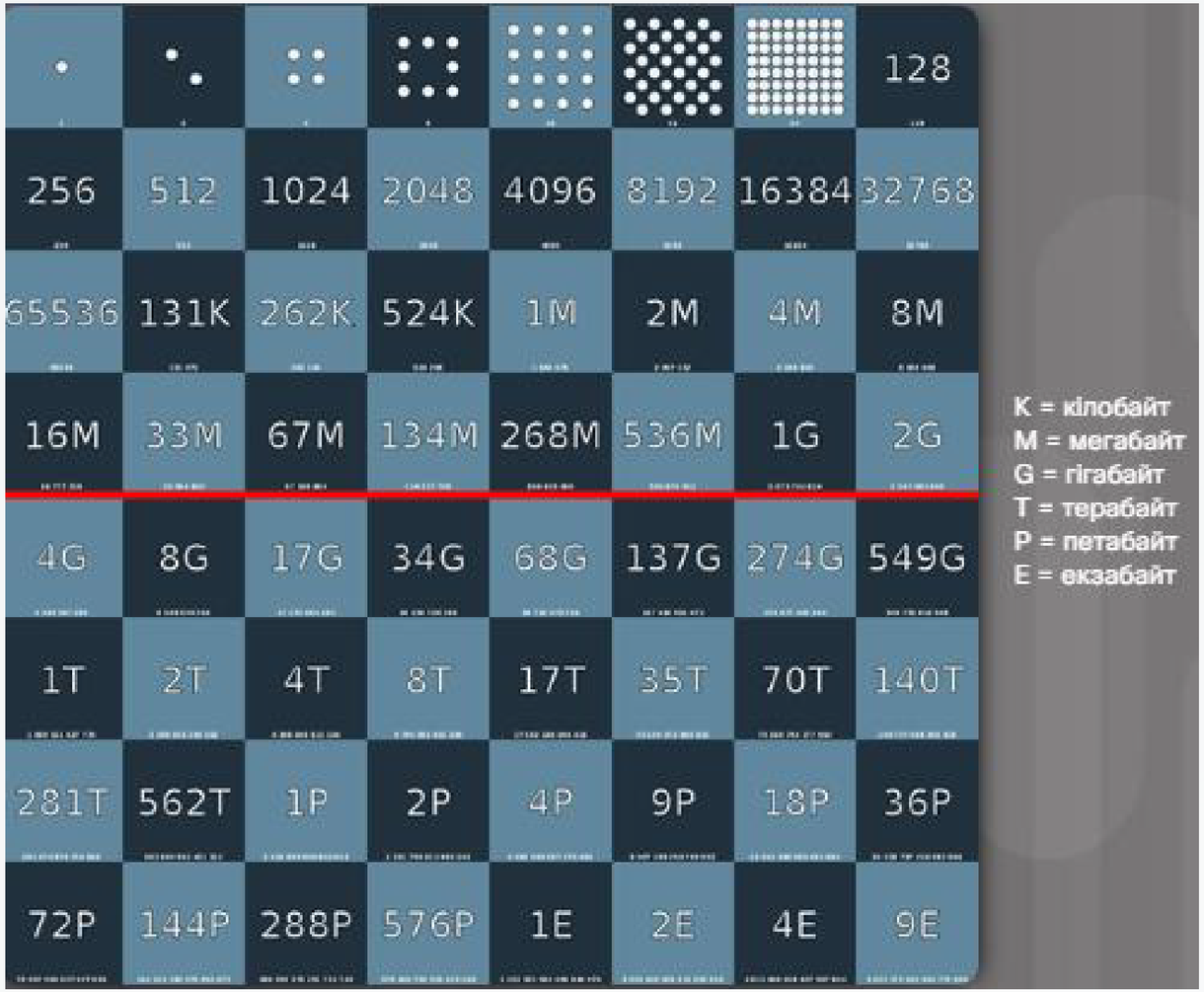
07 ПРИВАТНІСТЬ ДАНИХ

AGENDA

1.1. ІНТЕРНЕТ РЕЧЕЙ ТА ЗРОСТАННЯ ДАНИХ

- Зростаюча складність світу: Сучасний світ є надзвичайно складним, що призводить до постійного зростання обсягу генерованих даних.
- Швидкість генерації даних: Генерація даних не припиняється, і швидкість її зростає, що створює великий обсяг інформації.
- Інтернет речей (IoT): Розвиток Інтернету речей призводить до з'єднання різних пристроїв і генерації величезного обсягу даних.
- Потреба в зберіганні та аналізі даних: Зростаючий обсяг даних ставить завдання збереження та аналізу даних у новому контексті.
- Важливість якості даних: Не всі зібрані дані є корисними, тому важливо очищення та обробка даних для їхнього використання.
- Аналіз даних та візуалізація: Аналіз даних допомагає виявити цікаві відомості та тенденції, що можуть призвести до нових запитів.
- Великі дані (Big Data): Збільшення обсягу даних може призвести до вражаючих вимірів, що варто розглядати та управляти.

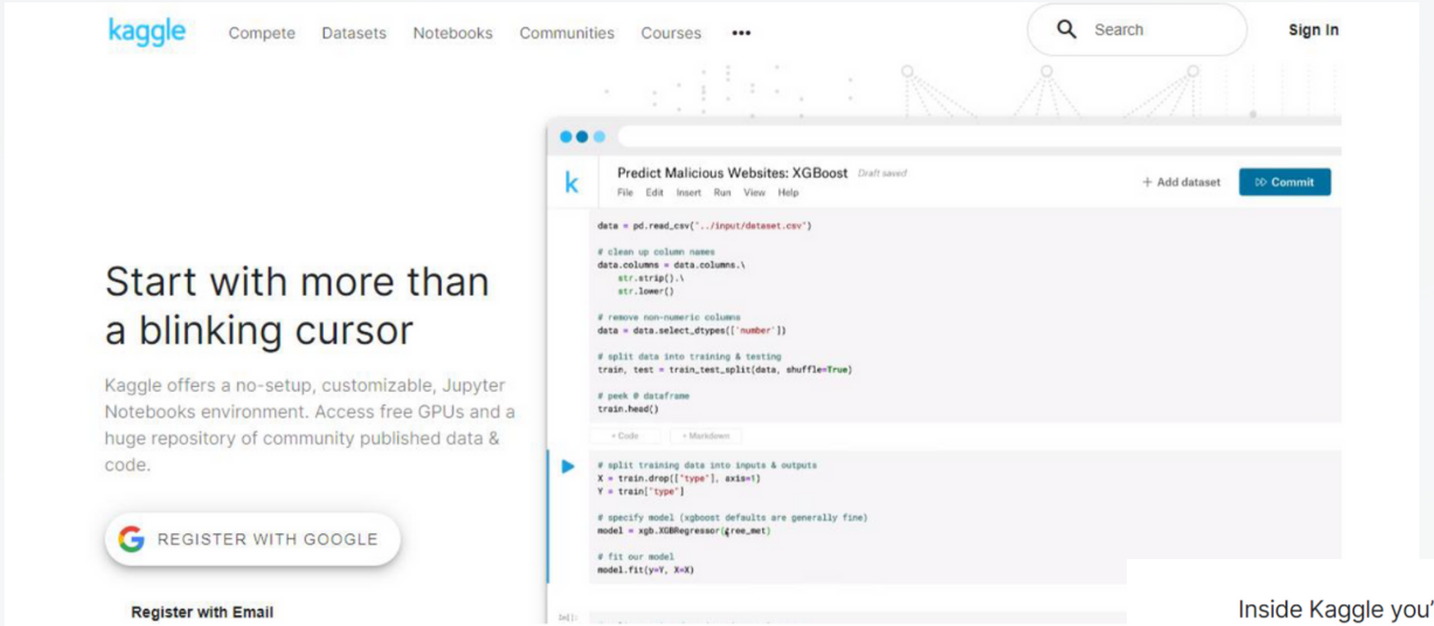
ПОДВОЄННЯ КІЛЬКОСТІ БАЙТІВ В КОМІРКАХ ШАХОВОЇ ДОШКИ



1.2. KAGGLE

- Набори даних: Kaggle має бібліотеку з тисячами наборів даних у різних галузях, що можуть бути використані для ваших проектів.
- Керування проектами: Ви можете створювати та керувати проектами, включаючи завдання, документацію та код.
- Kernel: Kernel – це інтерактивне середовище для роботи з даними та кодом. Ви можете використовувати Kernel для аналізу даних, побудови моделей машинного навчання та створення візуалізацій.
- Спільні ресурси: Kaggle дозволяє користувачам обмінюватися кодом та дослідженнями через публікації та обговорення.
- Конкурси: Kaggle відомий своїми конкурсами з машинного навчання, де учасники можуть змагатися за призи та розвивати свої навички.
- Навчання: Kaggle пропонує безліч безкоштовних ресурсів для навчання, включаючи курси, датасети та вправи.
- Документація та форуми: Крім того, на Kaggle є документація та форуми, де користувачі можуть знайти відповіді на свої запитання та отримувати допомогу.

KAGGLE



Inside Kaggle you'll find all the code & data you need to do your data science work. Use over 50,000 public datasets and 400,000 public notebooks to conquer any analysis in no time.

Machine Learning

Machine Learning is the hottest field in data science, and this track will get you started quickly

65k

Pandas

Short hands-on challenges to perfect your data manipulation skills

87k

Python

Learn the most important language for Data Science

65k

Deep Learning

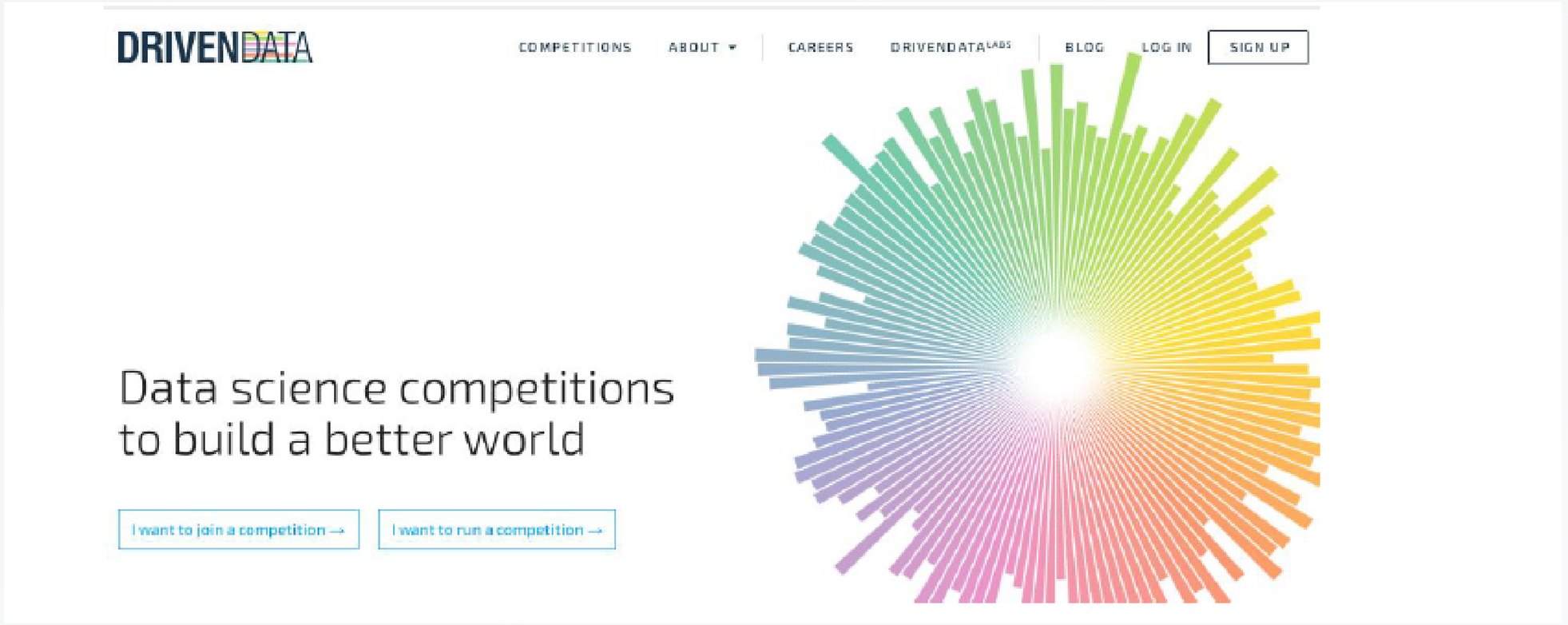
Use TensorFlow to take Machine Learning to the next level. Your new skills will amaze you

12k

1.3. DRIVENDATA

- Застосування IoT для збору даних: Використання технологій IoT дозволяє збирати велику кількість даних з різних джерел, включаючи датчики та сенсори, що допомагає вирішувати соціальні проблеми.
- Аналітика даних для прогнозування: Зібрані дані аналізуються для прогнозування різних тенденцій і розв'язання соціальних проблем, таких як утримання водяних насосів.
- Підвищення ефективності обслуговування: Прогнозування роботи пристроїв обслуговування за допомогою зібраних даних допомагає підприємствам створювати ефективніші плани обслуговування.
- Змагання для вирішення соціальних завдань: Платформа DrivenData організовує змагання, що залучають глобальну спільноту науковців для створення моделей, які можуть вирішити соціальні проблеми.
- Співпраця з некомерційними організаціями: DrivenData співпрацює з некомерційними організаціями, щоб розробити і впровадити найкращі моделі та інноваційні підходи для розв'язання соціальних завдань.

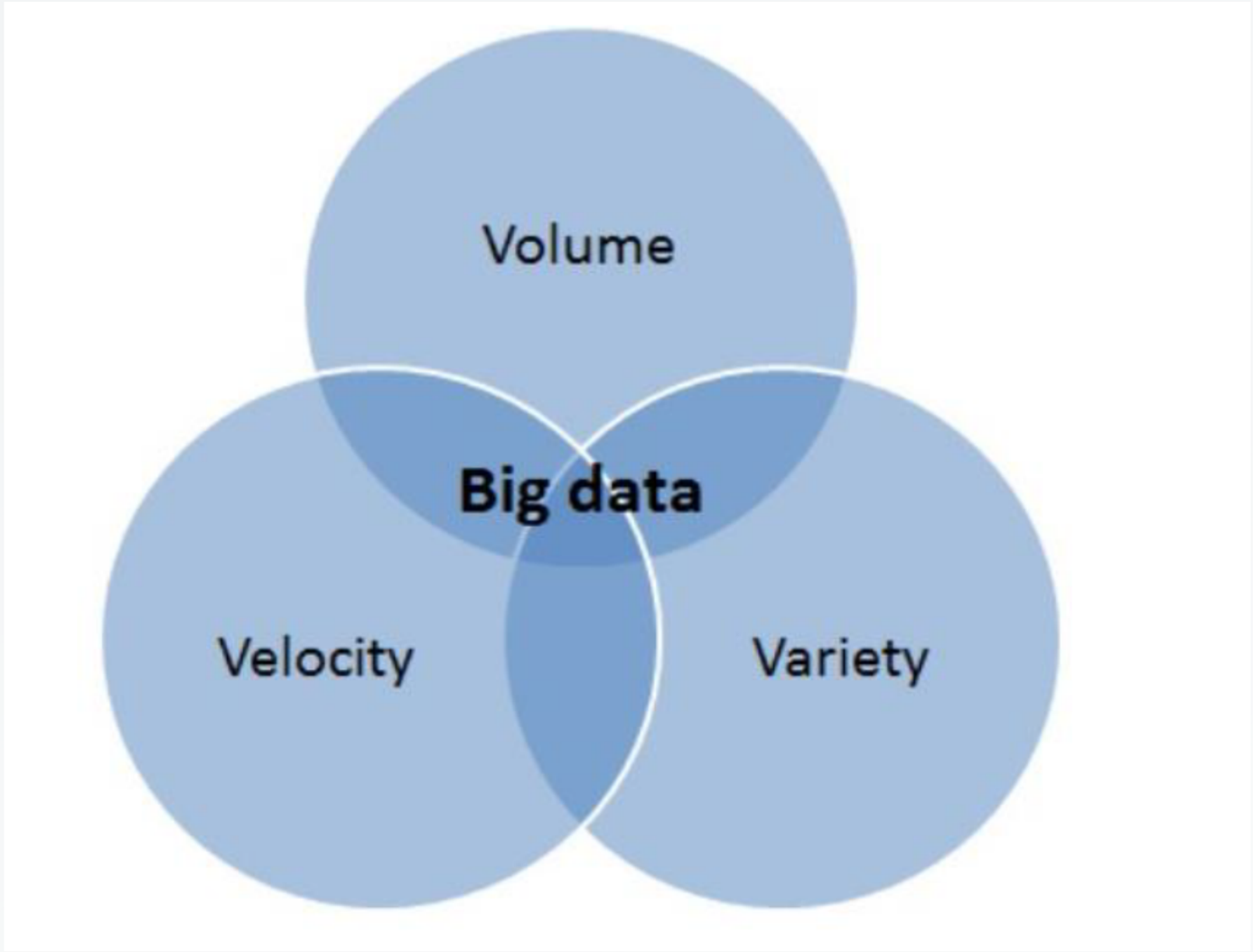
DRIVENDATA


DRIVENDATA
COMPETITIONS
ABOUT
CAREERS
DRIVENDATA LABS
BLOG
LOG IN
SIGN UP
Data science competitions
to build a better world
I want to join a competition →
I want to run a competition →

1.4. ВИЗНАЧЕННЯ ВЕЛИКИХ ДАНИХ

- Поняття Big Data: Big Data – це термін, що вказує на великі обсяги даних, які важко зберігати, обробляти та аналізувати за допомогою традиційних методів інформаційної обробки.
- Критерії Big Data: Об'єм, швидкість і різноманітність є ключовими критеріями, які визначають дані як Big Data. Обсяг може варіюватися, але важливо, щоб дані були настільки великими або складними, що вимагають спеціалізованих підходів до обробки.
- Виклики обробки Big Data: Одним з основних викликів у роботі з Big Data є потреба в обробці даних у режимі реального часу та інтеграції різних типів даних, включаючи неструктуровані дані.
- Роль технологій Big Data: Технології Big Data, такі як розподілені системи зберігання та обробки даних, грають важливу роль у вирішенні проблем, пов'язаних з Big Data.
- Значення точності та достовірності: Для використання Big Data важливо враховувати аспекти достовірності та точності даних, оскільки неправильні дані можуть призвести до недостовірних аналітичних висновків.

ХАРАКТЕРИСТИКИ BIG DATA



1.5. ПРИКЛАДИ ВЕЛИКИХ ДАНИХ У РЕАЛЬНОМУ СВІТІ

Генератори великих даних у реальному світі: Приклади, такі як Airbus A380 Engine, Large Hadron Collider та Square Kilometre Array, демонструють вражаючий обсяг генерації даних у різних галузях. Великі обсяги даних, які вони створюють, ставлять виклики перед сучасними технологіями обробки та аналізу даних.

Роль Інтернету речей (IoT): Застосування датчиків у сфері Інтернету речей приводить до надзвичайного зростання обсягів даних. Датчики, вбудовані у пристрої та обладнання, забезпечують постійний потік інформації, що сприяє експоненційному збільшенню Big Data.

Важливість обробки та аналізу великих даних: Завдяки генерації великих обсягів даних в реальному часі, обробка та аналіз даних стають ключовими завданнями для отримання цінної інформації. Використання спеціалізованих технологій та аналітичних підходів стає критичним для вивчення цих даних і прийняття стратегічних рішень.

1.6. ВІДКРИТІ ДАНІ

Роль відкритих даних в сучасному світі: Відкриті дані стають ключовим ресурсом для суспільства та бізнесу, сприяючи прозорості уряду, розвитку інновацій та розширенню можливостей аналізу та використання інформації.

Значення конфіденційності та приватності в аналітиці даних: Підвищення обсягу даних також вимагає посилення уваги до питань конфіденційності та захисту особистих даних. Забезпечення безпеки та приватності користувачів є важливим елементом в роботі з великими обсягами даних.

Приклади успішного використання відкритих даних: Сайти, як Портал відкритих даних Нью-Йорка та Gapminder, демонструють, як відкриті дані можуть бути використані для забезпечення доступу до інформації та сприяти розумінню соціальних та екологічних питань.

Потенціал відкритих даних в Україні: Портал відкритих даних України відкриває можливості для громадян, дослідників та бізнесу отримувати доступ до різноманітних наборів даних, що сприяє розвитку інновацій та покращенню рішень у різних сферах життя.

1.7. ПРИВАТНІСТЬ ДАНИХ

Революція в обробці даних: Постійний попит на дані для розробки програмних додатків та бізнес-рішень призводить до революції в обробці даних, де акцент зміщується з обміну необробленими даними на надання відповідей на конкретні запити користувачів.

Захист конфіденційності в епоху даних: З виникненням нових підходів до обробки даних, таких як SafeAnswers та openPDS, стає зрозумілим, що конфіденційність користувачів має стати пріоритетом, а не побічним ефектом використання даних.

Майбутнє конфіденційності та дотримання законодавства: Забезпечення конфіденційності даних користувачів вимагає більш широкого підходу, який охоплює не лише законодавство, але й зміну парадигми організаційної культури, де конфіденційність стає стандартним режимом роботи.

1.8. СТРУКТУРОВАНІ ТА НЕСТРУКТУРОВАНІ ДАНІ

Структуровані та неструктуровані дані: Відмінність між структурованими і неструктурованими даними полягає у способі їхнього упорядкування та доступності. Структуровані дані розміщуються у фіксованих полях і можуть бути легко оброблені за допомогою SQL та інших інструментів, тоді як неструктуровані дані не мають фіксованого формату і включають текст, фотографії, відео та інше.

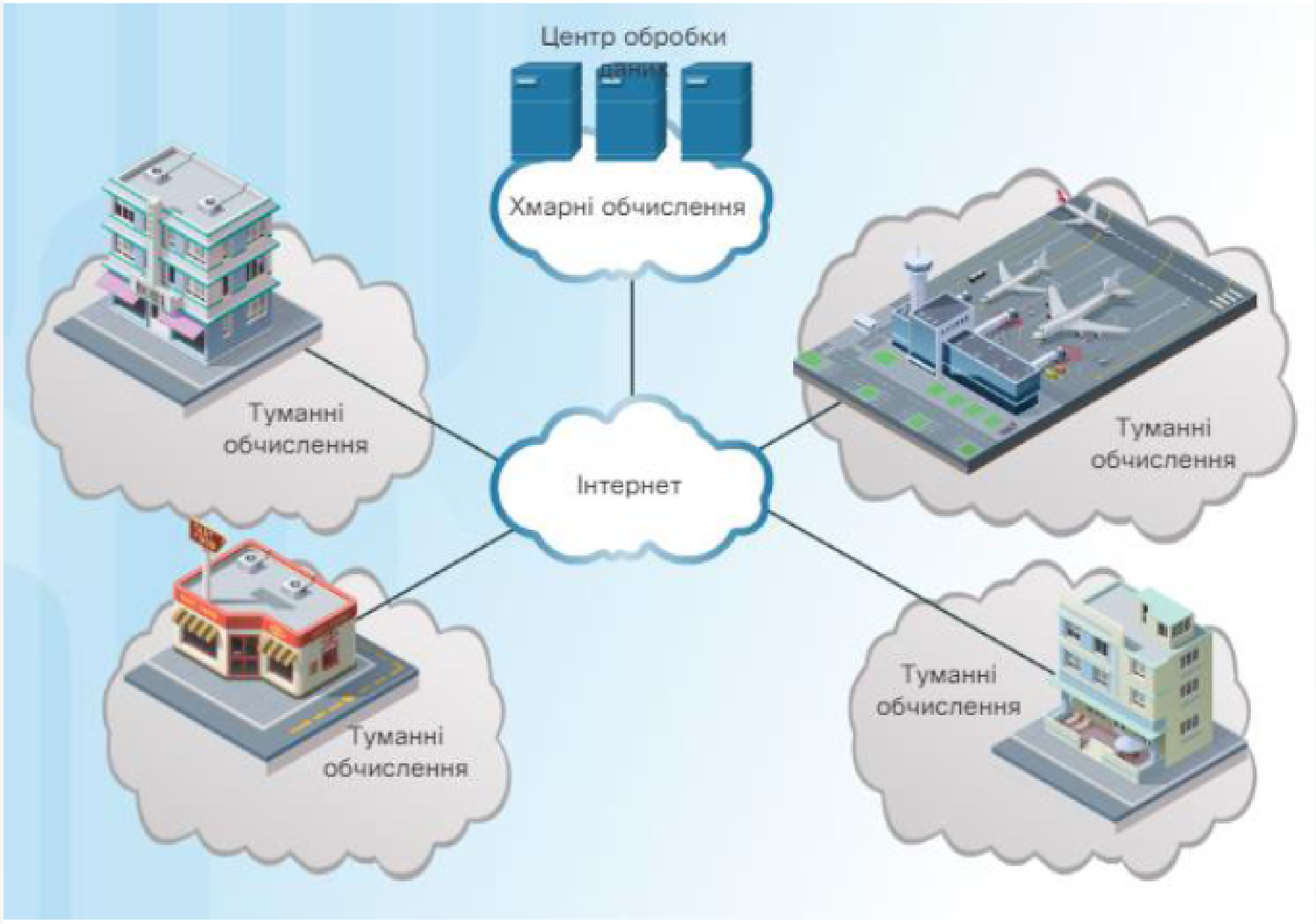
Значення неструктурованих даних: Неструктуровані дані становлять значну частину сучасних наборів даних, і вони мають великий потенціал для видобутку цінної інформації. Аналіз неструктурованих даних може призвести до виявлення нових тенденцій, патернів та можливостей для бізнесу та науки.

Необхідність управління різними типами даних: Організаціям важливо розробити стратегії для управління як структурованими, так і неструктурованими даними. Це включає в себе розробку методів форматування, зберігання та аналізу цих даних, щоб забезпечити їхню цінність і використання.

1.9. ХМАРНІ ТА ТУМАННІ ОБЧИСЛЕННЯ

- Зміна у способі зберігання та оброблення даних: Раніше дані зазвичай зберігалися на одному сервері або в локальних системах і оброблялися з використанням SQL. Зараз, з розвитком хмарних обчислень, дані можуть знаходитися в розподілених центрах обробки даних та бути доступними для аналізу ближче до їх джерела.
- Туманні обчислення як реакція на вимоги IoT: За появи Інтернету речей (IoT) зросла потреба у реальному часі аналізу та обробці великої кількості даних, що генеруються датчиками та пристроями. Туманні обчислення надають можливість обробляти дані ближче до джерела їх створення.
- Роль туманних обчислень в мережах IoT: Туманні обчислення сприяють оптимізації використання енергії, пропускної здатності мережі та зниженню затримок в обробці даних в мережах IoT. Вони дозволяють аналізувати дані на місці їхнього виникнення та передавати лише необхідну інформацію у хмару.
- Мінімізація затримок для реального часу: Затримки в мережі можуть стати перешкодою для аналізу даних в режимі реального часу. Використання туманних обчислень допомагає зменшити ці затримки та забезпечити оперативну обробку інформації.

МОДЕЛЬ ТУМАННИХ ОБЧИСЛЕНЬ



1.10. ДАНІ В СПОКОЇ ТА ДАНІ В РУСІ

- Дінамічні дані в русі та їх важливість: Динамічні дані, що перебувають у русі, вимагають оброблення в режимі реального часу та можуть призводити до негайних дій на основі аналізу цих даних. Це важливо для галузей, які полагаються на швидке реагування на зміни, наприклад, сільське господарство.
- Переваги обробки даних на межі: Обробка даних на межі, де вони створюються, дозволяє знизити обсяги даних, які потрібно зберігати централізовано, і дозволяє приймати рішення ближче до місця їх виникнення. Це особливо актуально для галузей, які мають велику кількість сенсорів і датчиків.
- Нові вимоги до зберігання та аналізу Big Data: Завдяки росту кількості датчиків та збільшенню їх оброблювальної потужності, дані можна аналізувати та використовувати ближче до їх джерела, зберігаючи лише необхідну інформацію. Ця зміна дозволяє отримувати більше цінної інформації в режимі реального часу і спонукає до швидших та більш точних рішень.

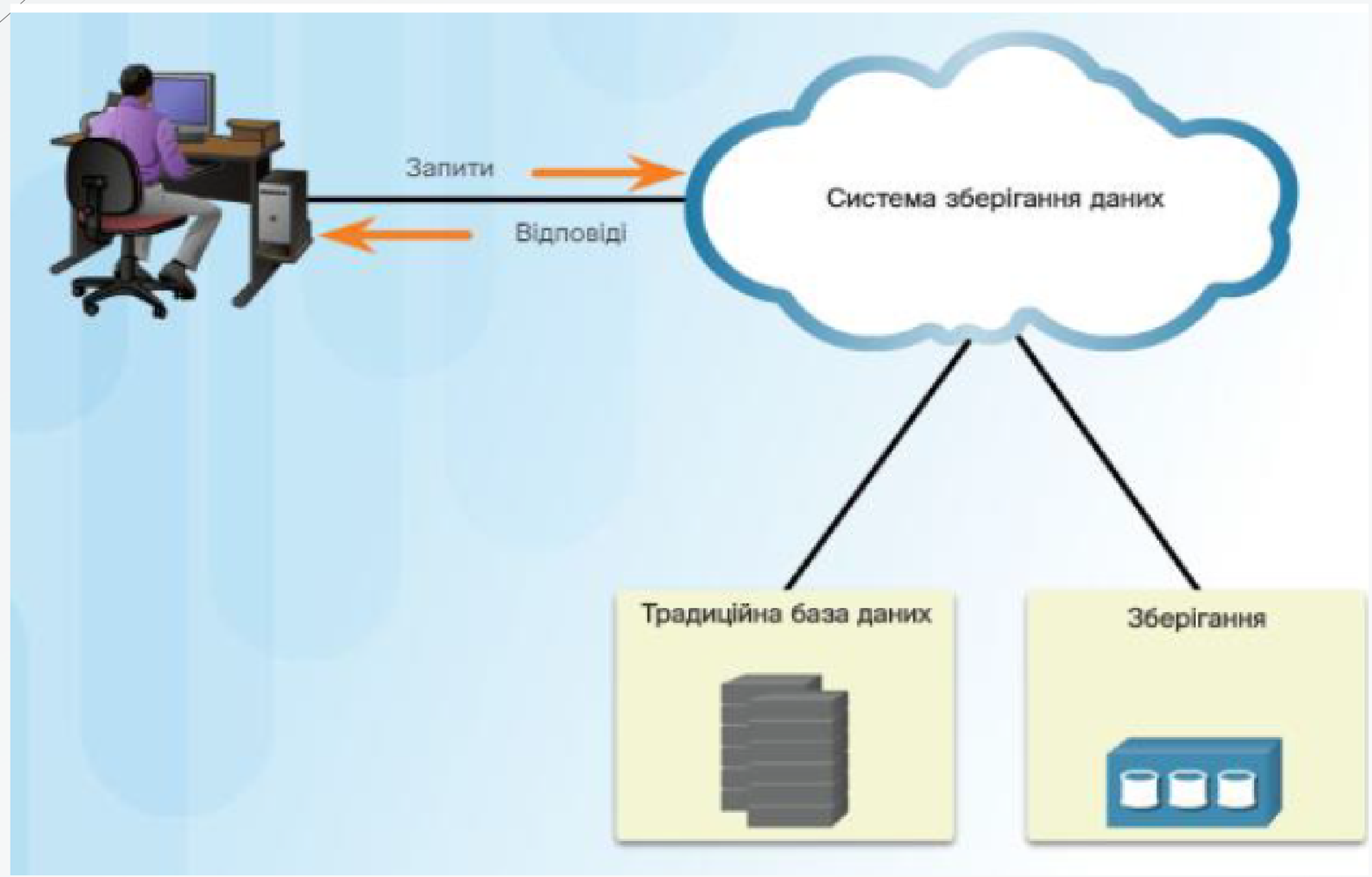
1.11. ІНФРАСТРУКТУРА ВЕЛИКИХ ДАНИХ

- Зміна інфраструктури для Big Data: Багато компаній переходять від традиційних серверів баз даних до розподіленої системи даних для забезпечення масштабованості та ефективної обробки великих наборів даних.

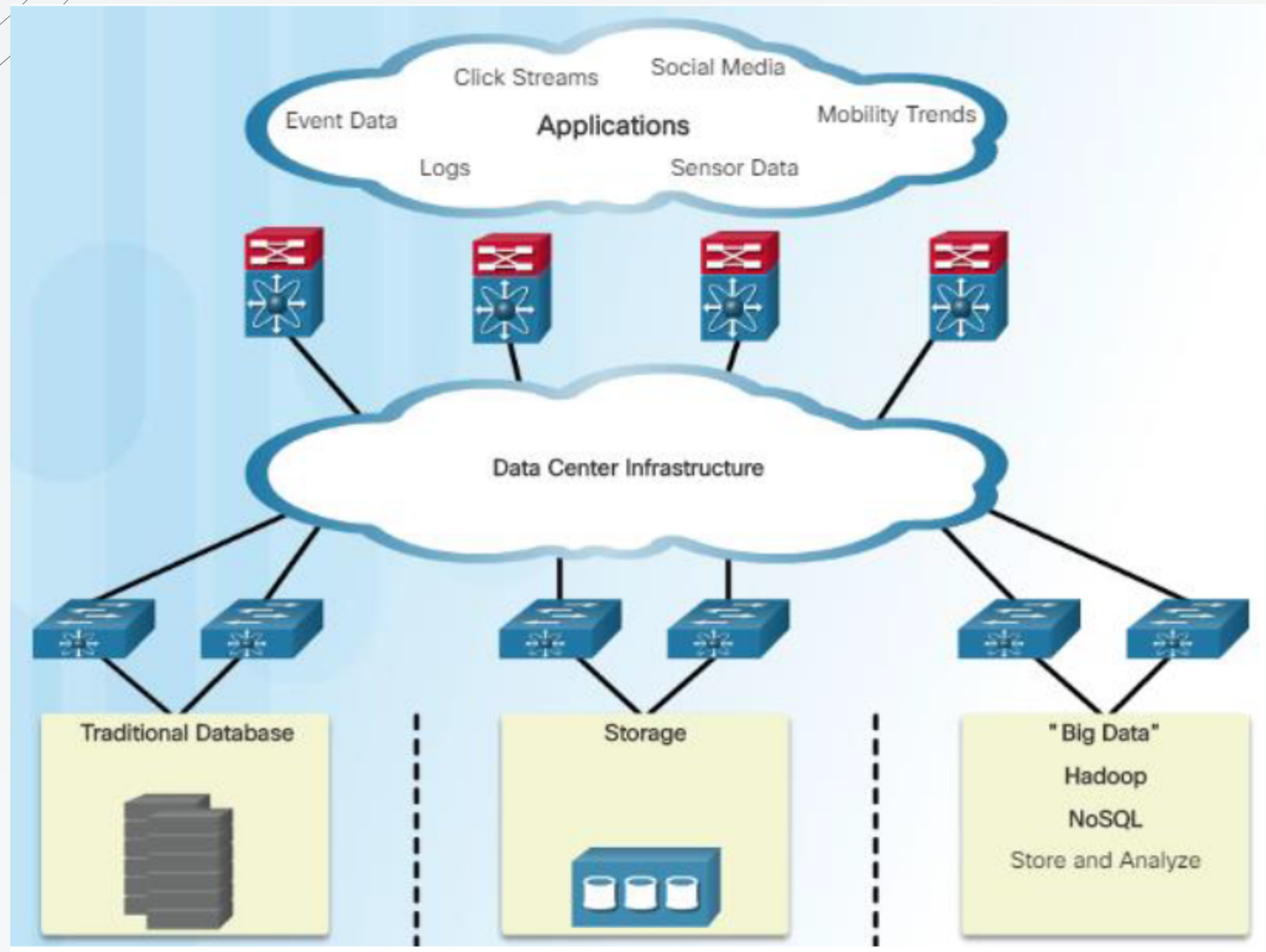
- Горизонтальна масштабованість: Горизонтальна масштабованість відрізняється від вертикальної масштабованості, оскільки вона не полягає в додаванні потужності до існуючих машин, а використовує розподілені ресурси для забезпечення доступу до даних багатьом користувачам одночасно.

- Значення доступу до даних: Забезпечення доступу до даних для багатьох користувачів одночасно стає все важливішим для компаній, оскільки вони використовують технології Big Data для керування бізнес–аналітикою та прийняття рішень.

ТРАДИЦІЙНА СИСТЕМА УПРАВЛІННЯ БАЗАМИ ДАНИХ



ІНФРАСТРУКТУРА ВЕЛИКИХ ДАНИХ



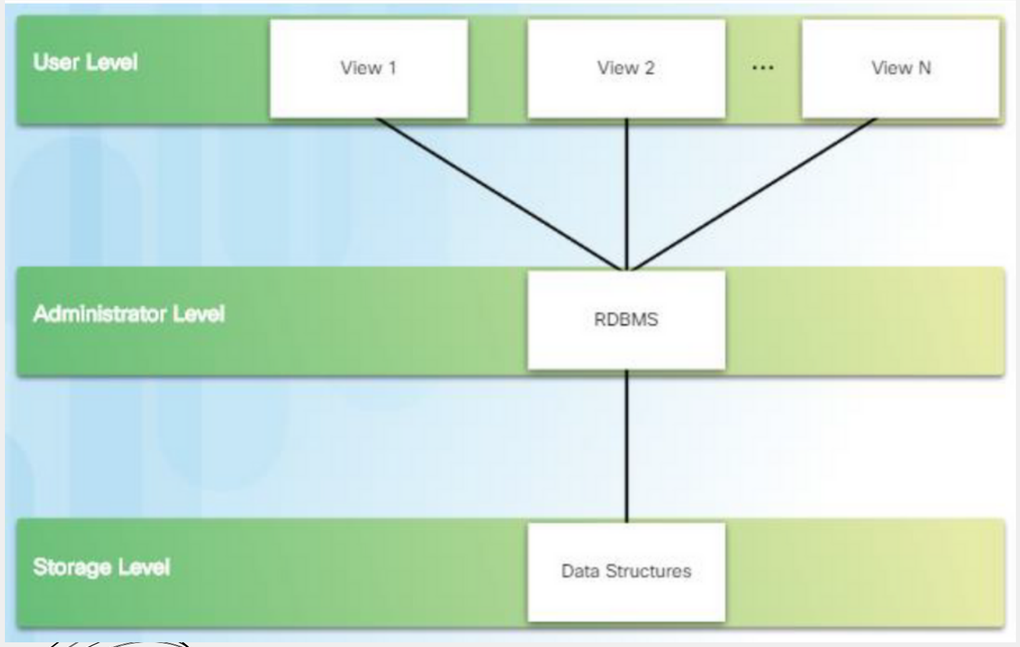
1.12. РОЗПОДІЛЕНІ ДАНІ ТА ЇХ ОБРОБКА

- Зміна управління даними з використанням RDBMS: Покоління управління даними рухається від використання систем управління реляційними базами даних (RDBMS) до нових підходів.
- Роль SQL в комерційних RDBMS: Багато комерційних рішень RDBMS використовують SQL як мову запитів для доступу до даних.
- Рівні абстракції управління даними: Управління даними включає низький рівень фізичного зберігання, опис збережених даних та рівень доступу користувачів до даних.
- Використання NoSQL для обробки великих даних: Бази даних NoSQL використовуються для обробки великих обсягів даних та веб-додатків у реальному часі, що дозволяє ефективно вирішувати бізнес-проблеми.
- Зростаючий обсяг даних та виклики аналітики: З поширеністю автоматизації бізнесу та вибуховим зростанням даних аналітиці стає важче керувати, особливо за допомогою традиційних RDBMS.

1.12. РОЗПОДІЛЕНІ ДАНІ ТА ЇХ ОБРОБКА

Розподілена обробка даних і Hadoop: Hadoop та інші розподілені системи допомагають обробляти великі обсяги даних шляхом розбиття їх на менші частини та обробки паралельно.

Використання мов програмування для аналізу даних: Для складних завдань аналізу даних використовуються мови програмування, такі як R та Python, які дозволяють створювати потужні інструменти аналізу даних.



АБСТРАГУВАННЯ ДАНИХ У РЕЛЯЦІЙНІЙ БАЗІ ДАНИХ