**ЗАТВЕРДЖЕНО**

Науково-методичною радою Державного університету «Житомирська політехніка»

протокол від \_\_ \_\_\_\_\_\_\_ 20\_\_ р. №\_\_

**КОНСПЕКТ ЛЕКЦІЙ**

**з навчальної дисципліни**

**«Аналіз великих даних у фінансах»**

для здобувачів вищої освіти освітнього ступеня «магістр»

спеціальності 051 «Економіка»

освітньо-професійна програма «Економіка»

факультет бізнесу та сфери обслуговування

(назва факультету)

кафедра цифрової економіки та міжнародних економічних відносин

(назва кафедри)

 (назва кафедри)

Рекомендовано на засіданні кафедри цифрової економіки та міжнародних економічних відносин

 (назва кафедри)

\_\_\_ \_\_\_\_\_\_\_\_\_\_ 2021 р., протокол № \_\_\_

 **Розробник**

Бондарчук Віталій Вікторович - заступник завідувача кафедри, доцент кафедри міжнародних економічних відносин, кандидат економічних наук

The Big Data Analytics in finance syllabus is developed in the framework of ERASMUS+ CBHE project «Digitalization of economic as an element of sustainable development of Ukraine and Tajikistan» / DigEco618270-EPP-1-2020-1-LT-EPPKA2-CBHE-JP

*This project has been funded with support from the European Commission. This document reflects the views only of the author, and the Commission cannot be held responsible for any use which maybe made of the information contained therein.*

*Цей проект фінансується за підтримки Європейської Комісії. Цей документ відображає лише погляди автора, і Комісія не несе відповідальності за будь-яке використання інформації, що міститься в документі.*

 Житомир

2021-2022 навчальний рік

Аналіз великих даних у фінансах

[Електронний ресурс] : конспект лекцій з дисципліни «Аналіз великих даних у фінансах» для здобувачів вищої освіти «Магістр» за спеціальністю: 051 «Економіка». – Житомир : Державний університет «Житомирська політехніка», 2021. – Режим доступу:

Конспект лекцій розроблено в рамках проєкту Erasmus+ “Діджиталізація економіки як елемент сталого розвитку України та Таджикистану (DigEco) 618270-EPP-1-2020-1-LT-EPPKA2-CBHE-JP”/ The proramm is developed in the framework of ERASMUS+ CBHE project “Digitalization of economic as an element of sustainable development of Ukraine and Tajikistan” / DigEco 618270-EPP-1-2020-1-LT-EPPKA2-CBHE-JP

Цей проект фінансується за підтримки Європейської Комісії. Цей документ відображає лише погляди автора, і Комісія не несе відповідальності за будь-яке використання інформації, що міститься в документі/This project has been funded with support from the European Commission. This document reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained there in.

Розробник: Бондарчук Віталій Вікторович - заступник завідувача кафедри, доцент кафедри міжнародних економічних відносин, кандидат економічних наук

© Державний університет «Житомирська політехніка», 2021

 © В. В. Бондарчук, 2021

**ЗМІСТ**

[**ТЕМА 1. ВСТУП. ПОНЯТТЯ ВЕЛИКИХ ДАНИХ ТА ЇХ КЛАСИФІКАЦІЯ. КОНЦЕПЦІЇ ВЕЛИКИХ ДАНИХ 4**](#_Toc82468680)

[**1. Суть Big Data 4**](#_Toc82468681)

[**2. Історія виникнення терміну Big Data 6**](#_Toc82468682)

[**3. Характеристики Big Data 7**](#_Toc82468683)

[**4. Основні типи Big Data 9**](#_Toc82468684)

[**ТЕМА 2. ПОНЯТТЯ РИНКУ ВЕЛИКИХ ДАНИХ. ЖИТТЄВИЙ ЦИКЛ АНАЛІТИКИ ДАНИХ. ЗБІР ТА ПІДГОТОВКА ДАНИХ 11**](#_Toc82468685)

[**1. Ринок великих даних: переваги, недоліки та ризики 11**](#_Toc82468686)

[**2. Поняття життєвого циклу великих даних 13**](#_Toc82468687)

[**3. Джерела даних 15**](#_Toc82468688)

[**4. Збір та підготовка даних 16**](#_Toc82468689)

[**ТЕМА 3. ТЕХНОЛОГІЇ ЗБЕРІГАННЯ ВЕЛИКИХ ДАНИХ 19**](#_Toc82468690)

[**1. Концепції зберігання великих даних 19**](#_Toc82468691)

[**2. Дискові системи зберігання даних 23**](#_Toc82468692)

[**3. Системи зберігання в оперативній пам'яті 32**](#_Toc82468693)

[**ТЕМА 4. ТЕХНОЛОГІЇ АНАЛІЗУ ДАНИХ 35**](#_Toc82468694)

[**1. Статистичний аналіз 36**](#_Toc82468695)

[**2. Машинне навчання 39**](#_Toc82468696)

[**3. Семантичний аналіз 44**](#_Toc82468697)

[**4. Візуальний аналіз 47**](#_Toc82468698)

[**ТЕМА 5. ВІЗУАЛІЗАЦІЯ ВЕЛИКИХ ДАНИХ 51**](#_Toc82468699)

[**1. Поняття візуалізації. Мова візуалізації 51**](#_Toc82468700)

[**2. Сфера застосування візуалізації та завдання які вона виконує 53**](#_Toc82468701)

[**3. Види візуалізації 54**](#_Toc82468702)

[**4. Коротка характеристика інструментів для візуалізації даних 58**](#_Toc82468703)

[**ТЕМА 6. ВЕЛИКІ ДАНІ ФІНАНСОВОЇ СФЕРИ 64**](#_Toc82468704)

[**1. Місце великих даних у фінансовій сфері 65**](#_Toc82468705)

[**2. Приклади використання великих даних в провідних країнах 67**](#_Toc82468706)

[**3. Використання великих даних в банках України 71**](#_Toc82468707)

[**ТЕМА 7. БЕЗПЕКА ТА КОНФІДЕНЦІЙНІСТЬ ВЕЛИКИХ ДАНИХ У ФІНАНСАХ 75**](#_Toc82468708)

[**1. Поняття інформаційної безпеки 75**](#_Toc82468709)

[**2. Підходи до безпеки на основі великих даних 77**](#_Toc82468710)

[**3. Безпека та конфіденційність даних у фінансовій сфері (на прикладі банківських установ) 81**](#_Toc82468711)

[**РЕКОМЕНДОВАНА ЛІТЕРАТУРА 87**](#_Toc82468712)

# ТЕМА 1. ВСТУП. ПОНЯТТЯ ВЕЛИКИХ ДАНИХ ТА ЇХ КЛАСИФІКАЦІЯ. КОНЦЕПЦІЇ ВЕЛИКИХ ДАНИХ

1. Суть Big Data

2. Історія виникнення терміну Big Data

3. Характеристики Big Data

4. Основні типи Big Data

*Конспект лекції укладено на основі джерел:*

Великі дані URL: https://uk.publicspeakingtip.org/big-data-7429

Кобзаренко Д.Н., Мустафаев А.Г. Учебное пособие дисциплины«Анализ больших данных» для направления подготовки 38.03.05«Бизнесинформатика», профиль «Электронный бизнес». – Махачкала: ДГУНХ, 2019 г.– 107 с.

Радченко И.А, Николаев И.Н. Технологии и инфраструктура Big Data. СПб:Университет ИТМО, 2018. 52 с.

Силен Дэвид, Майсман Арно, Али Мохамед Основы Data Science иBig Data. Python и наука о даных. СПб.: Питер, 2017. 336 с.

Томас Єрл, Ваджид Хаттак, Пол Булер Основи Big Data: Концепції, алгоритми та технології/Пер.з англ. Анатолія Гладуна;За наук.ред. Олексія Найди. Дніпро: «Баланс Бізнес Букс», 2018. 320 с.

**1. Суть Big Data**

Big Data (великі дані) – це поєднання структурованих, напівструктурованих та неструктурованих даних, які можуть бути видобуті для отримання інформації та використані в проектах машинного навчання, прогнозного моделювання та інших передових програм аналітики.

Системи, які обробляють і зберігають Big Data, стали загальним компонентом архітектур управління даними в великих організаціях.

Компанії використовують накопичені в їх системах Big Data для поліпшення операцій, забезпечення кращого обслуговування споживачів, створення персоналізованих маркетингових кампаній на основі конкретних уподобань клієнтів і, зрештою, підвищення прибутковості.

Підприємства, які використовують великі дані, мають потенційну конкурентну перевагу перед тими, хто цього не робить. Вони можуть приймати швидші та більш обґрунтовані ділові рішення, за умови, що вони ефективно використовують дані.

Наприклад, Big Data можуть надати компаніям цінну інформацію про своїх клієнтів. Вона може бути використана для вдосконалення маркетингових кампаній з метою збільшення залучення клієнтів та коефіцієнтів конверсії.

Крім того, використання великих даних дозволяє компаніям дедалі краще орієнтуватися на споживача.

Історичні дані та дані в реальному часі можуть бути використані для оцінки мінливих уподобань споживачів. Це дозволить підприємствам оновлювати та вдосконалювати свої маркетингові стратегії та ставати більш чутливими до бажань та потреб клієнтів.

Великі дані також використовуються медичними дослідниками для виявлення факторів ризику захворювання та лікарями для діагностики захворювань та станів у окремих пацієнтів.

Крім того, дані, отримані з електронних медичних записів, соціальних мереж, Інтернету та інших джерел, надають організаціям охорони здоров’я та державним установам найсвіжішу інформацію про загрози інфекційних захворювань чи спалахи захворювання.

В енергетичній галузі Big Data допомагають нафтогазовим компаніям визначати потенційні місця буріння та контролювати експлуатацію трубопроводі. Так само комунальні служби використовують їх для спостереженням за електричними мережами.

Фірми фінансових послуг використовують системи Big Data для управління ризиками та аналізу ринкових даних у реальному часі.

Виробники та транспортні компанії покладаються на великі дані для управління своїми ланцюгами поставок та оптимізації шляхів доставки.

Інші сфери використання включають – реагування на надзвичайні ситуації, запобігання злочинності та побудова розумних міст.

Великі дані надходять з безлічі різних джерел, таких як системи ділових операцій, бази даних клієнтів, медичні записи, журнали кліків в Інтернеті, мобільні додатки, соціальні мережі, сховища наукових досліджень, машинно генеровані дані та датчики даних в реальному часі, що використовуються в Інтернеті речей.

Дані можуть залишатися в необробленому вигляді в системах великих даних або попередньо оброблятися за допомогою інструментів інтелектуального аналізу даних або програмного забезпечення для того, щоб вони стали готові до конкретного використання в аналітиці.

Можна навести наступні приклади Big Data:

*Порівняльний аналіз.*Включає вивчення показників поведінки користувачів та спостереження за діями клієнтів у реальному часі з метою порівняння продуктів, послуг та авторитету однієї компанії з продуктами її конкурентів.

*Відстеження соціальних мереж.*Це інформація про те, що люди говорять у соціальних мережах про конкретний бізнес чи товар. Ці дані можуть бути використані, щоб допомогти визначити цільову аудиторію для маркетингових кампаній.

*Маркетинговий аналіз.*Сюди входить інформація, яка може бути використана для просування нових продуктів, послуг та ініціатив.

*Аналіз задоволеності споживачів та їх настроїв.*Вся зібрана інформація може показати, як клієнти ставляться до компанії чи бренду, як можна зберегти їх лояльність до бренду та як покращити зусилля щодо обслуговування клієнтів.

**2. Історія виникнення терміну Big Data**

В 2011 рокці поняття Big Data почало набирати популярність, в основному, у великих корпораціях таких як Microsoft, IBM, Oracle, EMC, HP та інших.

В 2011 році компанія Gartner відмітила великі дані як тренд номер два в інформаційно-технологічній інфраструктурі після віртуалізації. За прогнозами мається на увазі, що впровадження технологій Big Data суттєво вплине на інформаційні технології в сферах виробництва, охороні здоров'я, торгівлі, державного управлінні, а також в галузях, в яких реєструються індивідуальні переміщення ресурсів. З 2013 року Big Data починають викладати в університетах в рамках вузівських програм з науки про дані і інженерії.

Інноваційні розробки в області Big Data починалися не в маленьких стартапах, як це часто буває в IT-індустрії, а в великих компаніях. Так, наприклад, технологія розподіленої обробки даних MapReduce була розроблена компанією Google, a Hadoop, що є вільним програмним забезпеченням для виконання розподілених обчислень на кластерах з сотень і тисяч вузлів, відразу після створення активно підтримала компанія Yahoo. Більшість програмних продуктів в області Big Data є вільними, а їх адаптацією і просуванням займаються ті самі стартапи. Традиційні постачальники рішень в області зберігання і обробки даних, такі як IBM уважно ставляться до нових розробок в області Великих Даних і намагаються використовувати їх в своїх продуктах разом зі своїми технологіями.

**3. Характеристики Big Data**

Існує безліч характеристик для великих даних, але спробуємо розглянути основні.

Сфера великих характеризується такими ознаками:

*Volume (об'єм):* накопичена база даних є гігантський обсяг інформації, для якого обробка і зберігання традиційними способами є трудомісткими процесами. Такий обсяг потребує нових підходів і в більш вдосконалених інструментах.

*Velocity (швидкість):* цей показник вказує як на зростаючу швидкість накопичення, так і на швидкість обробки даних.

У багатьох випадках набори великих даних оновлюються в режимі майже реального часу, замість щоденних, щотижневих або щомісячних оновлень, характерних багатьом традиційним сховищам даних.

Програми аналітики великих даних співвідносять та аналізують вхідні дані, а потім надають відповідь або результат на основі запиту. Це означає, що аналітики даних повинні детально розуміти наявні дані та мати певне розуміння того, які відповіді вони шукають, щоб переконатися, що отримана інформація є дійсною та актуальною.

Управління швидкістю передачі даних також має важливе значення, оскільки аналіз великих даних поширюється на такі сфери, як машинне навчання та штучний інтелект, де аналітичні процеси автоматично знаходять закономірності у зібраних даних та використовують їх для отримання знань.

Останнім часом стали більш затребувані технології обробки даних в реальному часі.

*Variety (різноманіття):* дана характеристика означає можливість одночасної обробки структурованої і неструктурованої інформації різних форматів. Головною відмінністю структурованої інформації є можливість класифікації. Прикладом такої інформації може служити інформація про клієнтських транзакцій.

*Veracity (достовірність даних):* в даний час достовірність наявних даних є найважливішим критерієм для користувачів. Недостовірна інформація призводить до утруднення аналізу даних.

Достовірність даних стосується ступеня визначеності в наборах даних.

Невизначені необроблені дані, зібрані з різних джерел, таких як платформи соціальних медіа та веб-сторінки, можуть спричинити серйозні проблеми з якістю даних.

Наприклад, компанія, яка збирає масиви великих даних із сотень джерел, може виявити неточні дані, але аналітикам потрібна інформація про шляхи надходження даних, щоб простежити, де дані зберігаються, щоб вони могли виправити проблеми.

Неякісні дані призводять до неточного аналізу та можуть підірвати цінність бізнес-аналітики, оскільки це може призвести до недовіри керівників до даних у цілому.

Кількість невизначених даних в організації повинна бути врахована перед тим, як їх використовувати для аналізу великих даних. Командам ІТ та аналітики також потрібно забезпечити наявність достатньо точних даних для отримання достовірних результатів.

*Value (цінність накопиченої інформації):* великі дані повинні бути корисні в удосконаленні бізнес-процесів, складанні звітності або оптимізації витрат компаній.

Дуже важливо, щоб організації застосовували такі практики, як очищення даних, і існував механізм підтвердження, що дані стосуються відповідних питань бізнесу, перш ніж використовувати їх у проекті аналізу великих даних.

Перші три характеристики визначають так званий принцип «Трьох V».

Вирішальну роль у великих даних відіграють обсяг інформації, швидкість обробки, а також різноманітність з'являються даних.

*Обсяг* відноситься до наборів даних, розмір яких виходить за межі можливостей програмних засобів типової бази даних збору, зберігання, обробки і аналізу даних.

*Різноманітність* визначає здатність обробки безлічі типів, джерел і форматів даних від сенсорів, розумних пристроїв, соціальних мереж. Також різноманітність характеризується здатністю інтегрувати все більше число джерел, що містять різні структуровані, напівструктуровані дані, вилучаються з web-сторінок, web log файлів, e-mail, документів та ін.

*Швидкість* визначає реакцію на поточну інформацію за час, обмежене додатком. Прикладом є потокова обробка (наприклад, GPS даних в реальному часі).

**4. Основні типи Big Data**

Існує два типи даних – традиційні дані та великі дані.

Традиційні дані зберігаються в базах даних, які містять структуровані таблиці з текстовою, цифровою та іншою інформацією. Один комп’ютер може з легкістю управляти таким видом даних.

Традиційні дані можуть надходити з різних джерел. Як правило, це бувають дані про користувачів і клієнтів, наприклад, інформація про слухачів курсів з Data Science: повне ім’я, адреса, контактна інформація, кількість відвідувань або звернень до сервісного центру та ін.

У свою чергу, *великі дані* набагато перевершують в кількості традиційні дані. Такий тип даних розподіляється між комп’ютерами, але big data дуже важко використовувати ефективно. Ми отримуємо великі дані з абсолютно різних джерел – соціальних мереж (Facebook, Twitter, LinkedIn, Quora і т.д), фінансів, мобільних телефонів, курсів та інших ресурсів.

Big Data також охоплюють широкий спектр типів даних, включаючи наступні:

- структуровані дані в базах даних та сховищах даних на основі мови структурованих запитів (SQL);

- неструктуровані дані, такі як текстові та файли документів, що зберігаються в кластерах Hadoop або системах баз даних NoSQL

- напівструктуровані дані, такі як журнали веб-сервера або потокові дані з датчиків.

Всі різні типи даних можна зберігати разом за допомогою технологій які, як правило, базуються на Hadoop або службі зберігання хмарних об’єктів.

Крім того, програми для Big Data часто містять кілька джерел даних, які в іншому випадку не можуть бути інтегровані.

# ТЕМА 2. ПОНЯТТЯ РИНКУ ВЕЛИКИХ ДАНИХ. ЖИТТЄВИЙ ЦИКЛ АНАЛІТИКИ ДАНИХ. ЗБІР ТА ПІДГОТОВКА ДАНИХ

1. Ринок великих даних: переваги, недоліки та ризики

2. Поняття життєвого циклу великих даних

3. Джерела даних

4. Збір та підготовка даних

*Конспект лекції укладено на основі джерел:*

Силен Дэвид, Майсман Арно, Али Мохамед Основы Data Science иBig Data. Python и наука о даных. СПб.: Питер, 2017. 336 с.

Томас Єрл, Ваджид Хаттак, Пол Булер Основи Big Data: Концепції, алгоритми та технології/Пер.з англ. Анатолія Гладуна;За наук.ред. Олексія Найди. Дніпро: «Баланс Бізнес Букс», 2018. 320 с.

Фрэнкс БиллУкрощение больших данных: как извлекать знания из массивов информации с помощьюглубокой аналитики / Билл Фрэнкс ; пер. с англ. Андрея Баранова. М. : Манн, Иванов иФербер, 2014

Шандрівська О. Є., Кириленко А. А.Особливості ідентифікації ризиків ринку big data. URL: http://science.lpnu.ua/sites/default/files/journal-paper/2021/jun/23774/menedzhment121-84-97.pdf

**1. Ринок великих даних: переваги, недоліки та ризики**

Існуюча ринкова ситуація вимагає посилення орієнтації соціально-економічного розвитку окремих держав в частині забезпечення збалансування процесів трансформації національних ринків окремих країн з позиції поліпшення їх конкурентних позицій завдяки становлення у них цифрової економіки. Прикладні аспекти цифрової трансформації суспільства базуються на використанні технологій великих даних. Останні стають інструментом стратегічного планування, підвищення операційної ефективності, рівнів маркетингово-логістичного сервісу клієнтів в багатьох провідних компаніях. Підвищення точності прогнозування попиту споживачів, моделювання та візуалізація у процесі створення моделей нових продуктів і послуг, підтримка прийняття рішень, управління маркетинговими та логістичними ризиками, підвищення маржі на етапах створення доданої вартості тощо – лише деякі можливості системи інформаційно-аналітичного забезпечення підприємств на засадах використання масивів великих даних та цифрової обробки інформації. Очікується, що підвищення адаптаційної здатності завдяки роботі з великими даними, розвиток технологій захисту інформації та діджиталізація процесів сприятиме підвищенню інформаційної безпеки та запобіганню кіберзагрозам.

Ключовими гравцями на ринку великих даних є Китай, США, Канада, Франція та Великобританія. Серед великих учасників ринку великих даних є також суб’єкти регіону APAC: Індія, Південна Корея та Японія.

Значну частку глобального ринку Big Data становлять персональні дані фізичних осіб. В Україні деякі підприємства практикують систему обміну персональних даних клієнтів на персоналізовані послуги чи винагороди від підприємств, також є створені онлайнплатформи для опитувань (наприклад, Opinion.com.ua), за участь в яких користувачі отримують винагороду у вигляді віртуальних коштів, які, при накопиченні до певного обсягу, можливо конвертувати у реальні.

Попри те, що користувачі у більшості випадків погоджуються надавати свої персональні дані певним організаціям, вони є стурбованими щодо захисту персональної інформації цими організаціями. Споживачів хвилює ймовірність поширення організаціями їх персональних даних третім сторонам, а також ризик витоку даних через недостатньо потужну систему інформаційної безпеки підприємств.

Загалом можна виділити наступні сильні сторони ринку великих даних:

1) розширення бізнесу внаслідок збільшення обсягу інформації, якою він володіє ;

2) зростання кількості відгуків клієнтів через соціальні мережі;

3) встановлення стратегічних партнерських стосунків з постачальниками, дилерами та іншими зацікавленими особами завдяки застосуванню великих даних;

4) перманентне підвищення кваліфікації працівників для утримання конкурентоспроможності організацій;

5) налагоджена ІТ-система підприємства сприяє більш швидкому прийняттю ефективних управлінських рішень;

6) високі доходи, зумовлені прийняттям ефективних управлінських рішень, володіння результатами дослідження ринку завдяки великих даних;

7) здійснення прогнозування високої точності та визначення потенційних ризиків на основі використання великих масивів даних.

Не дивлячись на значні переваги, які обіцяють технології засновані на аналізі великих даних вони несуть, даний ринок має ряд загроз і є досить ризикованим, до таких ризиків можна віднести

1. Ризик втрати даних внаслідок хакерських атак. Ймовірність даного ризику зростає при збільшенні обсягу даних підприємства.

2. Ризик знищення конфіденційності даних.

3. Ризик зростання витрат на збір, обробку та зберігання даних. Помилка у плануванні бюджету може призвести до спіральних витрат, що у майбутньому спричинить анулювання доданої вартості, створеної завдяки використанню великих даних.

4. Ризик проведення неефективної аналітики зібраних даних.

5. Ризик збору неправдивих, некоректних, неякісних даних. Велика частка проектів є невдалими через використання неактуальних, застарілих або помилкових даних.

6. Ризик невідповідності дій над даними чинному законодавству.

7. Ризик формування висновків із низькою точністю.

8. Ризик порушення інтелектуальної власності третьої сторони.

9. Ризик виникнення етичних дилем.

10. Ризик хибної організації (структуризації) зібраних даних.

**2. Поняття життєвого циклу великих даних**

Аналіз великих даних відрізняється від традиційного аналізу даних в першу чергу з огляду на характеристику оброблюваних даних, таких як об'єм, швидкість і різноманітність. Для задоволення різних вимог до проведення аналізу великих даних необхідна поетапна методологія для організації дій і завдань, пов’язаних з придбанням, обробкою, аналізом і повторного використанням даних.

З точки зору впровадження великих даних і перспективного планування важливо, щоб крім життєвого циклу великих даних були враховані питання навчання, обладнання і кадрового забезпечення необхідного для аналітики даних.

Життєвий цикл аналітики великих даних можна розділити на дев'ять етапів:

1. Оцінювання бізнес-ситуації

2. Ідентифікація даних

3. Збір і фільтрація даних

4. Виокремлення даних

5. Перевірка і очищення даних

6. Агрегування і подання даних

7. Аналіз даних

8. Візуалізація даних

9. Використання результатів аналізу

Кожен життєвий цикл аналітики великих даних повинен виходити з чітко визначеної бізнес-ситуації,яка дає чітке уявлення про обґрунтування, мотивацію і цілі проведення аналізу. На етапах оцінювання бізнес-ситуації необхідно скласти економічне обґрунтування, оцінити і затвердити його до початку виконання реальних практичних завдань аналізу.

Оцінюючи бізнес-ситуацію потрібно чітко сформулювати мету для проведення аналізу великих даних, або іншими словами поставити проектне завдання.

Формування проектного завдання повинно включати такі моменти:

* Чітко сформульована мета досліджень.
* Призначення і контекст проекту.
* Попередній опис методики аналізу.
* Ресурси, які ви маєте намір використовувати.
* Доказ практичної можливості бути реалізованим проекту (або можливості перевірки концепції.)
* Пред'являються результати і критерій успіху.
* Календарний план.

На підставі отриманої інформації оцінюються витрати на проект, а також людські та інформаційні ресурси, необхідні для його успішного завершення.

**3. Джерела даних**

При зборі даних для подальшого аналізу важливим є визначитися з їх джерелами. Іноді потрібно збирати дані, як то кажуть, з нуля, але в багатьох випадках компанії вже мають певну базу даних, а іншу їх частину часто можна придбати у третіх сторін. Крім того слід мати на увазі що все більше організацій відкривають безкоштовний доступ до високоякісних даних для громадського і комерційного використання.

Перш за все оцінюють актуальність і якість даних, вже зібрані компанією. У багатьох компаніях існують спеціальні програми супроводу ключових даних, так що велика частина роботи з очищення даних може бути вже виконана. Ці дані можуть зберігатися в офіційних сховищах, якими керують ІТ-професіонали, але може бути і ситуація коли вони зберігаються в файлах Ехсе1 на комп'ютерах працівників компанії.

Знайти дані навіть в межах компанії може бути досить складно. В процесі зростання компанії її дані виявляються розсіяними по багатьом місцям. Дані можуть бути розкидані через те, що працівники переходять на інші посади або йдуть з компанії. Документація і метадані не завжди входять в число пріоритетів керівництва.

Отримання доступу до даних навіть в рамках окремо взятої компанії також може бути складаним завданням. Розуміють цінність і конфіденційність даних в компаніях досить часто встановлюються правила, при яких будь-якому працівнику доступні дані лише необхідні для його роботи. Ці правила перетворюються в фізичні і електронні бар'єри, які носять назву «китайські стіни». У більшості країн, в тому числі і в Україні, такі «стіни» щодо клієнтських даних є обов'язковими і суворо регламентованими.

Для проведення глибокого аналізу даних якими володіє компанія, як правило, недостатньо. Необхідні дані недоступні всередині організації, необхідно віднайти в зовнішньому світі. Багато компаній спеціалізуються на зборі цінної інформації. Наприклад, Nielsen та GFK добре відомі в цьому відношенні в сфері роздрібної торгівлі. Інші компанії надають дані для того, щоб ви, в свою чергу, удосконалювали надані ними послуги і екосистеми. Зокрема, до цієї категорії відносяться Twitter, Fasebook.

Хоча деякі компанії вважають дані дорогим ресурсом, в наші дні все більше урядових установ і організацій безкоштовно ділиться своїми даними. Це можуть бути досить якісні і правдиві дані в залежності від установи, яка створює їх і керує ними. Надана інформація відноситься до самих різних галузей. Інформація може принести користь як доповнення власних даних компаній, але вона також стане в нагоді тим, хто займається самонавчанням в галузі data science. Нижче в таблиці наведена невелика добірка постачальників відкритих даних, яких з кожним днем стає все більше і більше.

Постачальники відкритих даних

|  |  |
| --- | --- |
| Сайт з відкритими даними | Опис |
| Data.gov | Центр відкритих даних уряду США |
| https://open-data.europa.eu/ | Центр відкритих даних Європейської комісії |
| Data.worldbank.org | Проект відкритих даних всесвітнього банку |
| Data.gov.ua | Портал відкритих даних Міністерства цифрової трансформації України |

**4. Збір та підготовка даних**

Етап ідентифікації даних, присвячений визначенню наборів даних, необхідних для аналітичних проектів і їх джерел.

Залучення більш широкого спектра джерел даних може збільшити ймовірність виявлення прихованих закономірностей і кореляцій. Наприклад, щоб дати аналітичне висновок, може бити корисно визначити якомога більше типів пов'язаних джерел даних, особливо коли неясно, що саме потрібно шукати.

На етапах збору і фільтрації даних, вони збираються з усіх джерел, які були попередньо ідентифіковані. Потім отримані дані піддаються автоматизованій фільтрації для видалення пошкоджених або таких, що не мають особою значення для цілей аналізу.

Залежно від типу джерела дані можуть надходитияк набір файлів, наприклад, дані, отримані у стороннього постачальника, або можуть вимагати інтеграціїАРІ, наприклад, з Twitter. У багатьох випадках деякі або й більшість отриманих даних можуть бути нерелевантними і можуть бути відкинуті в процесі фільтрації.

Дані, що класифіковані як «спотворені», можутьвключати записи з відсутніми або безглуздимизначеннями або неприпустимо типами даних. дані,відфільтровані для одного аналізу, можуть бити значимість для іншого типу аналізу. Тому рекомендуєтьсязберегти точну копію вихідного набору даних передпочатком фільтрації.

Необхідно зберегти як внутрішні, так і зовнішні даніпісля генерування або використання всередині компанії.Для пакетної аналітики ці дані зберігаються на диску перед початком аналізу. У разі аналітики в реальному часі дані спочатку аналізуються, а потім зберігаються на диску.

Деякі дані, ідентифіковані як вхідні дані для аналізу, можуть надходити в форматі, несумісному для роботи з великими даними. Необхідність звертатися до несумісних типів даних більш імовірна при роботі з даними із зовнішніх джерел.

Необхідна ступінь виокремлення і перетворення залежать від типів аналітики і можливостей вирішення для великих даних.

Неправильні дані можуть спотворювати і фальсифікувати результати аналізу. На відміну від традиційних корпоративних даних, де структура даних заздалегідь визначена і дані попередньо перевірений, дані вводяться в аналіз великих даних можуть бити неструктурованих, без будь-яких вказівок на достовірність. Ця складність також може ускладнити отримання набору відповідних обмежень перевірки.

Етап перевірки і очищення даних призначений для створення складних правил перевірки і видалення любих відомих неприпустимо даних.

Рішення для великих даних часто отримують надлишкові дані в різних наборах даних. Ця надмірність може використовуватися для дослідження взаємопов'язаних наборів даних, щоб збирати параметри перевірки і заповнювати відсутні достовірні дані.

Для пакетної аналітики перевірка даних і їх очищення можуть бути виконані за допомогою автономної операції ETL.

Для аналітики в реальному часі потрібно більш складнасистема внутрішньої пам'яті для перевірки і очищення данихпо мірі їх надходження з джерела. Походження може відігравати важливу роль у визначенні точності і якості сумнівних даних. Дані, які здаються неприпустимими, можуть як і раніше мати значимість, оскільки вони можуть приховувати закономірності і тенденції.

Дані можуть бути розподілені за кількома наборами даних, вимагаючи об'єднання наборів даних через загальні поля, наприклад дату або ідентифікатор (ID). В інших випадках одні й ті ж поля даних можуть відображатися в декількох наборах даних, таких як дата народження.

Етап агрегування і представлення даних призначений для інтеграції декількох наборів даних разом для досягнення уніфікованого представлення.

Великі обсяги, оброблювані рішеннями для великих даних, можуть зробити агрегування даних довготривалим і трудомістким. Узгодження цих відмінностей може зажадати складної логіки, яка виконується автоматично без втручання людини.

# ТЕМА 3. ТЕХНОЛОГІЇ ЗБЕРІГАННЯ ВЕЛИКИХ ДАНИХ

1. Концепції зберігання великих даних

2. Дискові системи зберігання даних

3. Системи зберігання в оперативній пам'яті

*Конспект лекції укладено на основі джерел:*

Радченко И.А, Николаев И.Н. Технологии и инфраструктура Big Data. СПб:Университет ИТМО, 2018. 52 с.

Силен Дэвид, Майсман Арно, Али Мохамед Основы Data Science иBig Data. Python и наука о даных. СПб.: Питер, 2017. 336 с.

Томас Єрл, Ваджид Хаттак, Пол Булер Основи Big Data: Концепції, алгоритми та технології/Пер.з англ. Анатолія Гладуна;За наук.ред. Олексія Найди. Дніпро: «Баланс Бізнес Букс», 2018. 320 с.

**1. Концепції зберігання великих даних**

Через необхідність зберігання масивів даних великого об'єму були створені інноваційні стратегії зберігання і технології для досягнення економічних і масштабованих рішень по зберіганню даних. Для того, щоб зрозуміти основні механізми технології зберігання великих даних, розглянемо:

• кластери

• файлові системи та розподілені файлові системи

• NoSOL

• шардінг

• реплікація

• CAP теорема

• ACID

• BASE

**Кластери**

В даному випадку кластер є тісно пов'язаною сукупністю серверів або вузлів. Кожен вузол в кластері має свої власні ресурси, такі як пам'ять, процесор і жорсткий диск. Кластер може виконувати завдання, розбиваючи їх на дрібні частини і розподіляючи їх виконання на різних комп'ютерах, що належать кластеру.

**Файлові системи та розподілені файлові системи**

Файлова система - це спосіб зберігання і організації даних на запам'ятовуючих пристроях, таких як флеш-накопичувачі, DVD-диски і жорсткі диски. Файлова система забезпечує логічне представлення даних, що зберігаються на пристрої зберігання даних, і представляє його у вигляді дерева каталогів і файлів. Операційні системи використовують файлові системи для зберігання і зчитування даних за завданням додатків. Кожна операційна система підтримує одну або кілька файлових систем, наприклад NTFS в Microsoft Windows чи ext в Linux.

Розподілена файлова система – файлова система, яка може зберігати великі файли, розподілені по вузлах кластера.

Прикладами можуть бути файлова система Google (GFS) і розподілена файлова система Hadoop (HDFS).

**NoSQL**

База даних типу NoSQL є нереляційною базою даних, яка вирізняється високою масштабованістю, відмовостійкістю і розроблена спеціально для зберігання напівструктурованих і неструктурованих даних.

**Шардінг**

Шардінг - це процесі горизонтального поділу великого масиву даних на колекцію менших, які є більш керованими і називаються шардами. Шарди розподілені серед безлічі вузлів, де вузол - це сервер або комп'ютер. Кожен шард зберігається на окремому вузлі, і кожен вузол відповідальний тільки за дані, що зберігаються на ньому. Кожен сегмент спільно використовує одну і ту ж логічну структуру в базі даних, і все разом шарди є повним набором даних.

Шардінг дозволяє розподіляти навантаження по обробці між декількома вузлами для досягнення горизонтальної масштабованості.

Перевагою шардінгу є те, що воно забезпечує часткову відмовостійкість до збоїв. У разі збою вузла, тільки дані зберігаються на цьому вузлі схильні до впливу. Що стосується розбиття даних, то необхідно враховувати Шаблони запитів, щоб самі сегмент не ставали вузькими місцями продуктивності.

**Реплікація**

Реплікація зберігає безліч копій наборів даних, відомих як «репліки», на кількох вузлах.

Для реалізації реплікації використовуються два різних методи:

• ведучий-ведений

• одноранговий

**Режим реплікації «ведучий-ведений»**

Під час односпрямованого режиму реплікації «ведучий-ведений» вузли розміщуються в конфігурації ведучий-ведений, і всі дані записуються в головний вузол.

Після збереження даних реплікуються на кілька підлеглих вузлів. Всі зовнішні запити на запис, включаючи вставку, оновлення та видалення даних, відбуваються на головному вузлі, а запити на читання можуть виконуватися будь-яким підлеглим вузлом.

Режим односпрямованої реплікації «ведучий-ведений» ідеально підходить для читання інтенсивних завантажень більшою мірою, ніж для запису інтенсивних завантажень, оскільки зростаючі вимоги до читання можуть управлятися за допомогою горизонтального масштабування через додавання додаткових ведених вузлів.

Записи є узгодженими, оскільки всі записи координуються головним вузлом. Це означає, що продуктивність записів буде погіршуватися із збільшенням обсягу записів. Якщо ведучий вузол вийде з ладу, то читання буде можливим через будь-який з відомих вузлів.

**Режим реплікації «одноранговий»**

При реплікації в режимі «одноранговий» все вузли працюють на одному рівні. Кожен вузол, відомий як спеціальний робочий вузол, в рівній мірі здатний обробляти операції читання і запису. Кожен запис копіюється в усі однорангові вузли.

Тимчасова реплікація схильна до помилок записи, які виникають в результаті одночасного оновлення одних і тих же даних декількома одноранговими вузлами. Цю проблему можна вирішити шляхом реалізації стратегії розпаралелювання.

**Шардінг і реплікація**

Для поліпшення відмовостійкості, що забезпечується за допомогою використання прийомів шардінга, а також для отримання додаткових переваг від підвищеної доступності і масштабованості реплікації, можна комбінувати як шардінг, так і реплікацію ,.

Можливі наступні комбінації:

• шардінг і реплікація в режимі «ведучий-ведений»

• шардінг і реплікація в режимі «одноранговий»

Коли шардінг об'єднується з реплікацією в режимі «ведучий-ведений», то кілька шардів стають відомими одного ведучого, а сам ведучий також є шардом.

Коли шардінг об'єднується з реплікацією в режимі «одноранговий», кожен шард реплікується декільком одноранговим вузліам, і кожен одноранговий вузол відповідає тільки за підрядним із загального масиву даних. У сукупності це допомагає підвищити масштабованість і відмовостійкість.

**Теорема CAP**

Теорема про узгодженість, доступності та стійкості при поділі (CAP), також відома як теорема Брюєра, висловлює потрійне обмеження, пов'язане з розподіленими системами баз даних. Вона стверджує, що розподілена система баз даних, що працює в кластері, може забезпечити тільки два з наступних трьох властивостей:

• Узгодженість - читання даних з будь-якого вузла призводить до таких же даних, що знаходяться на декількох вузлах.

• Доступність. - Запит на читання / запис завжди буде підтверджений у вигляді успіху або відмови.

• Стійкість при поділі. - Система бази даних може бути відмовостійкої при збоях обміну даними в процесі поділу кластера на кілька сховищ даних і при цьому може обслуговувати запити читання / запису.

**ACID**

ACID - це принцип проектування бази даних, пов'язаний з управлінням транзакціями.

ACID - це стиль управління транзакціями, який задіює песимістичні елементи управління паралелізмом доступу для забезпечення узгодженості, підтримуваної шляхом застосування блокувань записів. ACID - це традиційний підхід до управління транзакціями бази даних, так як він використовується системами керування базами даних.

**BASE**

BASE є принципом проектування баз даних на основі теореми CAP і максимального використання концепцій систем баз даних, які використовують розподілену технологію. BASE розшифровується як: абсолютно доступні; м'який стан; випадкова узгодженість.

Коли база даних підтримує BASE, то це сприяє доступності завдяки узгодженості. По суті, BASE використовує оптимістичні елементи управління паралелізмом, послаблюючи сильні обмеження узгодженості, що визначаються властивостями ACID.

BASE-сумісні бази даних не є корисними для транзакційних систем, де відсутність узгодженості є проблемою.

**2. Дискові системи зберігання даних**

Дисковий сховище зазвичай використовує недорогі жорсткі диски для довготривалого зберігання даних. Дисковий сховище може бути реалізовано за допомогою розподіленої файлової системи або бази даних

Розподілені файлові системи, як і будь-яка файлова система, що не залежать від даних, що зберігаються і тому підтримують зберігання даних без схеми. У загальному випадку розподілена файлова система системи зберігання даних забезпечує вбудовану надлишковість і високу доступність шляхом копіювання даних в кілька локальних місць розташування за допомогою реплікації.

Розподілена файлова система не підходить для наборів даних з великою кількістю дрібних файлів, так як призводить до активізації надмірної пошукової активності на диску, тим самим сповільнюючи загальний доступ до даних.

Розподілена файлова система підходить для зберігання великих наборів необроблених даних або для архівування наборів даних. Крім того, вона забезпечує можливість недорогого зберігання великих обсягів даних, які повинні залишатися доступними в мережі (online), протягом тривалого періоду часу.

Слід зазначити, що розподілені файлові системи не пропонують можливість пошуку по вмісту файлів в якості стандартної вбудованої можливості.

**СУБД**

Системи управління базами даних (СУБД) підходять для обробки транзакційних робочих завдань з використанням невеликих обсягів даних з можливістю довільного читання / запису. СУБД є ACID-сумісними, і, дотримуючись це відповідність, вони, як правило, обмежені одним вузлом. З цієї причини СУБД не дають вбудованої надмірності та відмовостійкості.

Для обробки великих обсягів даних, що надходять в швидкому темпі, реляційні бази даних, як правило, потребують масштабування. У СУБД використовується вертикальне масштабування, а не горизонтальне, яке є більш дорогою і революційною стратегією масштабування. Але це робить СУБД менш придатними для довгострокового зберігання даних, які згодом накопичуються.

Реляційні бази даних, як правило, вимагають даних для прив'язки до схеми. В результаті зберігання слабоструктурованих і неструктурованих даних, схеми яких не є реляційними безпосередньо не підтримується. Це призводить до накладних витрат, які створюють затримку. Ця затримка робить реляційні бази даних не придатними для зберігання високошвидкісних даних, для який потрібне високонадійне обладнання зберігання баз даних з високою швидкістю запису.

В результаті через недоліки, традиційна СУБД зазвичай не використовується в якості основної системи зберігання в середовищі рішень для великих даних.

**Бази даних NoSQL**

Бази даних NoSQL (Not-only SQL) відносяться до технологій, які використовуються для розробки нереляційних баз даних наступного покоління, що володіють високою масштабованість і відмовостійкість.

Нижче наведено список основних властивостей систем зберігання даних NoSQL, которие відрізняють їх від стандартних СУБД.

• Модель даних без схеми - дані можуть існувати в необробленій формі.

• Горизонтальне масштабування – для отримання додаткового сховища з базою даних NoSQL можна додати додаткові вузли, замість заміни існуючого вузла більш досконалим, з кращими характеристиками / більшою місткістю.

• Висока доступність- будується на основі кластерних технологій, які забезпечують вбудовану відмовостійкість.

• Зниження експлуатаційних витрат - багато баз даних NoSQL побудовані на платформах з відкритим вихідним кодом без будь-яких витрат на ліцензування. Тому вони можуть бути розгорнуті на стандартному апаратному забезпеченні.

• Потенційна узгодженість - дані читаються на декількох вузлах, але можуть бути неузгодженими відразу після запису. Однак всі вузли будуть в кінцевому підсумку в узгодженому стані.

• ***BASE, без ACID —*** BASE-відповідність вимагає, щоб база даних підтримувала високу доступність в разі збою мережі / вузла, при цьому не обов'язково, щоб база даних знаходилася в узгодженому стані при кожному оновлень. База даних може бути в гнучкому / неузгоджену стані, поки вона в кінцевому підсумку не досягне узгодженості.

• Доступ до даних за допомогою АРІ - доступ до даних, як правило, підтримується за допомогою запитів на основі АРІ, включаючи API REST.

• Автоматичний шардінг і реплікація - для підтримки горизонтального масштабування і забезпечення високої доступні в системах зберігання даних NoSQL автоматично використовуються методи шардінгу і реплікації, коли набір даних розбивається горизонтально, а потім копіюється на кілька вузлів.

• Інтегроване кешування - усуває необхідність в сторонньому рівні розподіленого кешування, такому як Memcached.

• Розподілена підтримка запитів - системи зберігання даних NoSQL підтримує узгоджену поведінку запитів в декількох шардах.

• Орієнтування на агрегування - системи зберігання NoSQL зберігають денормалізовані агреговані дані, тим самим усуваючи необхідність в об'єднанні і масштабному зіставленні між об'єктами прикладного рівня і даними, що зберігаються в базі даних.

Системи зберігання даних бази NoSQL в основному можна розділити на чотири типи залежно від способу зберігання даних:

• ключ-значення

• документ

• об'єднання стовпців

• граф

*Ключ-значення*

Системи зберігання даних типу «ключ-значення» зберігають дані як пару з ключа і значення і діють як хеш-таблиці. Таблиця є списком значень, де кожне значення ідентифікується ключем. Це значення зазвичай зберігається як BLOB.

Збережене значення може бути будь-якою сукупністю, починаючи від даних датчика і закінчуючи відео.

Пошук значення може бути виконаний тільки за допомогою ключів, так як база даних не враховує деталі збереженої сукупності. Часткове оновлення неможливо. Оновлення є операцію видалення або вставки.

Зазвичай система зберігання «ключ-значення» не підтримує ніяких індексів, тому записи виконуються досить швидко. Ґрунтуючись на простій моделі зберігання, системи зберігання «ключ-значення» є високомасштабовані.

Оскільки ключі є єдиним засобом отримання інформації, зазвичай ключ додається до типу зберігається значення для легкого вилучення даних. Прикладом є 123\_sensorl.

Системи зберігання «ключ-значення» підходить для випадків, коли:

• потрібно сховище для неструктурованих даних

• потрібна висока ефективність читання / запису

• значення повністю ідентифікується тільки одним ключем

• значення - автономний об'єкт, що не залежить від інших значень

• значення мають порівняно просту структуру або є бінарними

• шаблони запитів прості, і включають в себе тільки операції вставки, вибору і видалення

• до збережених даних звертаються на прикладному рівні

Системи зберігання даних «ключ-значення» недоцільні, якщо:

• додаткам потрібно виконати пошук або фільтрацію даних з використанням атрибутів зберігається значення

• існують взаємозв'язки між різними записами пари «ключ-значення»

• значення групи ключів повинні оновлюватися в одній транзакції

• кілька ключів вимагають перетворення (звернення) за одну операцію

• потрібна узгодженість схеми для різних значень

• необхідне оновлення окремих атрибутів значення

Прикладами систем зберігання «ключ-значення» є Riak, Redis і Amazon Dynamo DB.

*Документ*

Системи зберігання «документів» також зберігають дані як пари з ключем і значенням. Однак, на відміну від систем зберігання «ключ-значення», збережене значення - це документ, на який можна зробити запит з бази даних. Ці документи можуть мати складну розгалужену структуру. Документи можна кодувати, використовуючи схему кодування на основі тексту, таку як XML або JSON, або використовуючи схему двійкового кодування, таку як BSON (Binary JSON).

Основні відмінності між системами зберігання «документів» і системами зберігання «ключ-значення» наступні:

• системи зберігання «документів» інформовані про значення

• збережене значення самостійно описується; схема може бути виведена зі структури значення або включена в значення посиланням на схему документа

• операція вибору може посилатися на поле всередині агрегованого значення

• операція вибору може отримати частину агрегованого значення

• підтримуються часткові поновлення; тому агреговане підмножина може бути оновлено

• як правило, підтримуються індекси, які прискорюють пошук

Кожен документ може мати альтернативну схему; тому можливо зберігання різних типів документів в одній і тій же колекції або блоці. Додаткові поля можуть додаватися в документ після початкової вставки, тим самим забезпечуючи гнучку підтримку схеми.

Система зберігання документів є доцільною, коли:

• зберігаються слабоструктуровані документо-орієнтовані дані, що містять плоску або розгалужену схему

• еволюція схеми є вимогою, через невідомість або ймовірності зміни структури документа

• додатки вимагають часткового оновлення агрегату даних, збереженого як документ

• пошукові запити повинні виконуватися в різних полях документів

• об'єкти домену, такі як клієнти, зберігаються в формі серіалізованого об'єкту

• шаблони запитів включають операції вставки, вибору, оновлення та видалення

Система зберігання документів є недоцільною, коли:

• необхідно оновити кілька документів в рамках однієї транзакцій

• виконуються операції, що вимагають об'єднання декількох документів або зберігання нормалізованих даних

• необхідне застосування схеми для доставки їх до обумовленого проектування запитів, так як структура документа може змінюватися між послідовними виконаннями запитів, що потребують реструктуризації запиту

• значення що зберігається не є самостійно описуваним і не має посилання на схему

• дані повинні зберігатися в двійковому коді

 Приклади систем зберігання документів включають MongoDB CouchDB і Terrastore.

*Об'єднання стовпців*

Системи зберігання даних типу «об'єднання стовпців» зберігають дані так само, як традиційні СУБД, але пов'язані стовпці згруповані в одному рядку, що призводить до утворення об'єднання стовпців. Кожен стовпець може бути колекцією пов'язаних стовпців.

Використовувати системи зберігання об'єднання стовпців доцільно, коли:

• необхідна можливість довільного читання / запису в реальному масштабі часу, і збережені дані мають певну структуру

• дані представляють собою табличну структуру, де кожен рядок складається з великого числа стовпців і існують вкладені групи взаємопов'язаних даних

• потрібна підтримка розвитку схеми, оскільки об'єднання стовпців можуть бути додані або видалені без будь-яких простоїв системи

• до деяких полях, найчастіше, звертаються разом, і пошук повинен виконуватися з використанням значень полів

• потрібне ефективне використання сховища даних, коли дані складаються з рідко заповнених рядків так як бази даних об'єднання стовпців розподіляють тільки простір пам'яті, якщо є в наявності стовпець для рядка. Якщо стовпця немає, простір не виділяється

• шаблони запитів включають в себе операції вставки, вибору, оновлення та видалення

Системи зберігання об'єднання стовпців недоцільні, коли:

• потрібно реляційний доступ до даних, наприклад, об'єднання

• потрібна підтримка ACID транзакцій

• двійкові дані повинні зберігатися

• повинні виконуватися SOL-сумісні запити

• шаблони запитів, можуть часто змінюватися, оскільки вони можуть ініціювати відповідну реструктуризацію розташування об'єднання стовпців

Приклади систем зберігання об'єднання стовпців включають в себе Cassandra, HBase і Amazon SimpleDB.

*Граф*

Графові системи зберігання використовуються для збереження взаємозалежних об'єктів. На відміну від інших систем зберігання NoSOL, де акцент робиться на структуру об'єктів, графові системи зберігання базуються на збереження зв'язків між об'єктами.

Об'єкти зберігаються у вигляді вузлів і називаються вершинами, в той час як зв'язки між об'єктами є ребрами. Мовою реляційних баз кожен вузол можна розглядати як рядок, а ребрами позначають зв'язок.

Вузли можуть мати більше одного типу зв'язків відображених декількома ребрами. Кожен вузол може містити дані атрибутів у вигляді пар «ключ-значення», наприклад, як вузол клієнта з атрибутами ідентифікатора особистості, імені та віку. Кожне ребро може мати свої власні дані атрибутів у вигляді пар «ключ-значення», які можуть використовуватися для подальшої фільтрації результатів запиту. Як правило, пристрої зберігання графів забезпечують узгодженість через відповідність ACID.

Ступінь корисності пристрої зберігання графів залежить від кількості і типів ребер, утворених між вузлами. Чим більше число ребер і їх різноманітність, тим різноманітніше типи запитів, які система може обробляти.

Графові системи зберігання зазвичай дозволяють додавати нові типи вузлів без внесення змін до бази даних.

Це також дозволяє визначати додаткові зв'язки між вузлами в міру появи нових типів відносин або вузлів в базі даних.

Графові системи зберігання доцільні, коли:

• необхідно зберігати взаємопов'язані об'єкти

• запит об'єктів ґрунтується на типі відносин між ними, а не на атрибутах об'єктів

• здійснюється пошук групи взаємопов'язаних об'єктів

• здійснюється пошук відстаней між об'єктами з точки зору відстані обходу вузла

• виконується інтелектуальний аналіз даних з метою пошуку закономірностей

Графові системи зберігання неприйнятні, якщо:

• необхідні оновлення для великого числа атрибутів вузлів або атрибутів ребер, оскільки це пов'язано з пошуком вузлів або ребер, що є дорогою операцією в порівнянні з виконанням обходів вузлів

• об'єкти мають велику кількість атрибутів або вкладених даних - найкраще зберігати легкі об'єкти в графових системах зберігання, зберігаючи інші дані атрибутів в окремій неграфової системі зберігання даних NoSQL

• потрібно сховище двійкових даних

• запити, засновані на вибори атрибутів вузла / ребра, домінують над запитами обходу вузлів

Приклади включають в себе Neo4J, Infinite Graph і OrientDB.

*Бази даних NewSQL*

Системи зберігання даних NoSQL є добре масштабованими, доступними, відказостійкими і швидкодіючими для операцій читання / запису.

Системи зберігання NewSQL поєднують в собі властивості ACID СУРБД з масштабністю і відказостійкістю, які пропонують системи зберігання NoSQL. Бази даних NewSQL підтримують SQL-совместімнй синтаксис для опису даних і операцій маніпулювання даними, а також вони часто використовують логічну реляційну модель даних для зберігання даних.

Бази даних NewSQL можуть використовуватися для розробки OLTP-систем з дуже великим об'ємом транзакцій, наприклад, для банківської системи. Вони також можуть використовуватися для аналітики в реальному масштабі часу, наприклад, оперативної аналітики.

У порівнянні з NoSQL, системи зберігання NewSQL забезпечує більш простий перехід від традиційних СУРБД до масштабованих баз даних завдяки підтримці SQL.

Прикладом баз даних NewSQL включають в себе VoltDB, NuoDB і InnoDB.

**3. Системи зберігання в оперативній пам'яті**

Системи зберігання в оперативній пам'яті зазвичай використовують ОЗП, головну пам'ять комп'ютера, в якості свого основного носія, щоб забезпечити швидкий доступ до даних. Збільшення ємності і зниження вартості ОЗП в поєднанні зі збільшенням швидкості читання / запису жорстких дисків, дозволило розробити ефективні рішення для зберігання даних в оперативній пам'яті.

Зберігання даних в оперативній пам'яті усуває затримку дискового введення / виводу і час передачі даних між основною пам'яттю і жорстким диском. Це загальне скорочення затримки читання / запису даних значно прискорює обробку даних.

Системи зберігання в оперативній пам'яті дозволяють виконувати аналітику в пам'яті, тобто шляхом виконання запитів до даних, що зберігаються в пам'яті, а не на диску. Виконання аналітики в оперативній пам'яті дозволяє прискорити операційну аналітику.

Оперативна пам'ять для великих даних реалізується в кластері, забезпечуючи високу доступність і надмірність даних. Таким чином, горизонтальної масштабованості можна досягти шляхом простого додавання великої кількості вузлів або пам'яті. Якщо порівнювати з дисковим пристроєм зберігання, оперативна пам'ять набагато дорожче через більш високу вартість пам'яті в порівнянні з дисковим пристроєм зберігання.

Оперативна пам'ять для великих даних реалізується в кластері, забезпечуючи високу доступність і надмірність даних. Таким чином, горизонтальної масштабованості можна досягти шляхом простого додавання великої кількості вузлів або пам'яті.

В першу чергу, оперативна пам'ять дозволяє відчути швидкий приплив даних в середовищі великих даних (характеристика швидкості), створюючи середовище зберігання даних, яке створює відчуття реального часу.

Залежно від реалізації, оперативна пам'ять може підтримувати зберігання без схеми або зберігання до схеми. Сховище без схеми здійснюється за допомогою збереження даних на основі «ключ-значення».

Оперативна пам'ять є актуальною, коли:

• дані надходять в швидкому темпі і вимагають аналітики в реальному часі або обробки потоку подій

• потрібна неперервна або постійна виконувана аналітика і операційна аналітика

• необхідно виконати інтерактивну обробку запитів і візуалізацію даних в реальному часі, включаючи аналіз даних «що-якщо» і операцій розгортки

• потрібно один і той же набір даних для декількох завдань обробки даних

• виконується аналіз даних, так як не потрібно перезавантажувати з диска один і той же набір даних, якщо алгоритм змінюється

• обробка даних передбачає ітераційний доступ до одного набору даних, наприклад, виконання алгоритмів, що базуються на графах

• розробка рішень для великих даних з малими затримками і підтримкою ACID транзакцій

Оперативна пам'ять не є актуальною, коли:

• обробка даних складається з пакетної обробки

• протягом тривалого часу необхідно зберігати дуже великі обсяги даних у внутрішній пам'яті для проведення глибокого аналізу даних

• виконується стратегічний ВІ або історичний аналіз, який передбачає доступ до дуже великих об’ємів даних і включає в себе пакетну обробку даних

• набори даних надзвичайно великі і не поміщаються в доступну пам'ять

• у разі переходу від традиційного аналізу даних до аналізу великих даних,

• підприємство має обмежений бюджет, так як установка оперативної пам'яті може вимагати від вас оновити вузлів, що можна зробити або шляхом заміни вузла, або шляхом додавання великої обсягу ОЗП

Пристрої оперативної пам'яті можуть реалізовуватися як:

• In-Memory Data Grid (IMDG)

• In-Memory Database (IMDB)

Хоча обидві ці технології використовують операційну пам'ять в якості основного носія даних, однак те, що робить їх відмінними, полягає в способі зберігання даних в пам'яті.

# ТЕМА 4. ТЕХНОЛОГІЇ АНАЛІЗУ ДАНИХ

1. Статистичний аналіз

2. Машинне навчання

3. Семантичний аналіз

4. Візуальний аналіз

*Конспект лекції укладено на основі джерел:*

Кобзаренко Д.Н., Мустафаев А.Г. Учебное пособие дисциплины«Анализ больших данных» для направления подготовки 38.03.05«Бизнесинформатика», профиль «Электронный бизнес». – Махачкала: ДГУНХ, 2019 г.– 107 с.

Радченко И.А, Николаев И.Н. Технологии и инфраструктура Big Data. СПб:Университет ИТМО, 2018. 52 с.

Силен Дэвид, Майсман Арно, Али Мохамед Основы Data Science иBig Data. Python и наука о даных. СПб.: Питер, 2017. 336 с.

Томас Єрл, Ваджид Хаттак, Пол Булер Основи Big Data: Концепції, алгоритми та технології/Пер.з англ. Анатолія Гладуна;За наук.ред. Олексія Найди. Дніпро: «Баланс Бізнес Букс», 2018. 320 с.

Ын Анналин, Су Кеннет Теоритический минимум по Big Data, что нужно знать о больших даных. СПб.: Питер, 2019. 208 с.

Аналіз великих даних поєднує традиційні методи статистичного аналізу даних з обчислювальними методами. Статистична вибірка із сукупності є ідеальною, коли доступний весь набір даних, і така умова типова для традиційних сценаріїв пакетної обробки. Однак великі дані можуть перейти від пакетної обробки до обробки в реальному часі через необхідність розуміти сенс потокових даних. У разі потокових даних набори даних накопичуються з плином часу, і дані упорядковуються в часі. У потокових даних акцент робиться на своєчасну обробку, так як аналітичні результати мають термін придатності. Незалежно від того, чи є це розпізнавання можливості продажу, яка випливає з поточного контексту клієнта, або виявленням аномальних умов у виробничому середовищі, що вимагають втручання для захисту обладнання або забезпечення якості продукти, істотне значення має час, а також актуальність аналітичного результати.

У будь-якому швидко мінливому просторі, такому як великі дані, завжди існують можливості для інновацій. Прикладом цього є питання про те, як найкращим чином поєднувати статистичні та обчислювальні підходи для даної аналітичної задачі. Статистичні методи широко використовуються для попереднього аналізу даних, після чого можуть застосовуватися обчислювальні методи, які використовують інформацію, витягнуту з результатів статистичного аналізу набору даних. Перехід від пакетної обробки до обробки в реальному часі представляє ще одну трудність, оскільки методи обробки в реальному часі повинні використовувати ефективні для обчислення алгоритми.

**1. Статистичний аналіз**

Статистичний аналіз використовує статистичні методи, які опираються на математичні формули як засіб для аналізу даних. Статистичний аналіз найчастіше є кількісно, але може бити і якісним.

Цей тип аналізу зазвичай використовується для опису наборів даних за допомогою підсумкового узагальнення, наприклад, надання середнього значення, медіани або методу статистики, пов'язаних з набором даних. Також він може використовуватися для визначення моделей і відносин всередині набору даних, таких як регресія і кореляція.

Розглянемо наступні види статистичного аналізу:

• А / В-тестування

• кореляція регресія

**А / В-тестування**

А / В-тестування, також відоме як спліт-тестування або тестування за допомогою «відра», порівнює дві версії елементів для визначення, яка з них є більш досконалою на основі визначеної метрики. Елементів може представляти цілий ряд речей. Наприклад, це може бути контент, такий як веб-сторінка або пропозицію продукту або послуги, наприклад, у вигляді ціни на електронні товари. Поточна версія елементів називається контрольної версією, тоді як змінена версія називається розрахунковою. Обидві версії одночасно піддаються експерименту. Спостереження записуються для визначення, яка версія є більш успішною.

Хоча А/В-тестування може бути реалізовано практично для будь-якої області, найчастіше воно використовується в маркетинг. Як правило, мета полягає в тому, щоб оцінювати поведінку людини для можливості збільшення продажів.

Приклади питань можуть включати в себе:

• Чи є новий препарат краще, ніж старий?

• Клієнти краще реагують на рекламні оголошення, відправлені електронною або звичайною поштою?

• Чи є новостворений дизайн головної сторінки Web-сайту джерелом генерування більшої кількості трафіку від користувачів?

**Кореляція**

Кореляція - це метод аналізу, який використовується для визначення, того, чи пов'язані дві змінні один з одним. Якщо виявиться, що вони пов'язані між собою, наступним кроком буде визначення їх відносини. Наприклад, значення змінної А збільшується щоразу, коли збільшується значення змінної В. Нас може зацікавити також визначення того, наскільки тісно пов'язані змінні А і В, а значить, ми також можемо проаналізувати, якою мірою змінна В збільшується по відношенню до збільшення змінної А.

Використання кореляції допомагає розвинути розуміння набору даних і знайти відносини, які можуть допомогти в поясненні явищ. Тому кореляція зазвичай використовується для Data Mining, де визначення взаємозв'язків між змінними в наборі даних призводить до виявлення закономірностей і аномалій. А це в свою чергу розкриває природу набору даних або причину явища.

Коли дві змінні вважаються корельованими, вони вирівнюються на основі лінійної залежності. Це означає, що коли одна змінна змінюється, інша змінна також змінюється пропорційно і постійно.

Кореляція виражається у вигляді десяткового числа від -1 до +1, яке називається коефіцієнтом кореляції. Ступінь відносини змінюється від сильної до слабкої при переході від -1 до 0 або +1 до 0.

Приклади питань, при кореляції:

• Чи впливає віддаленість від моря на температуру в місті?

• Учні, з хорошою успішність в молодшій школі, мають таку ж хорошу успішність в середній школі?

• Якою мірою ожиріння пов'язане з переїданням?

**Регресія**

Аналітичний метод регресії досліджує, як пов'язана залежна змінна з незалежною змінною в наборі даних. Наприклад, регресія може допомогти визначити тип відносин між температурою, незалежної змінної, і врожайністю культур, залежною змінною.

Застосування цього методу допомагає визначити, як значення залежної змінної змінюється щодо змін до значень незалежної змінної. Коли, наприклад, збільшується незалежна змінна, залежна змінна також збільшується? Якщо так, збільшення лінійне або нелінійне?

Одночасно можна протестувати кілька незалежних змінних. Однак в таких випадках змінюватися може тільки одна незалежна змінна, в той час як інші залишаються незмінними. Регресія може допомогти краще зрозуміти явище, і чому воно відбулося. А також вона може використовуватися для прогнозування значень залежної змінної.

Лінійна регресія являє собою постійну швидкість зміни.

Приклади питань можуть включати в себе:

• Яка буде температура в місті, що знаходиться в 250 милях від моря?

• Якими будуть оцінки учня, що вчиться в середній школі, ґрунтуючись на його оцінках в початковій школі?

• Які шанси, що людина буде страждати ожирінням, ґрунтуючись на кількості споживаної їжі?

Регресія і кореляція мають ряд важливих відмінностей. Кореляція не припускає причинності. Зміна значення однієї змінної може не відповідати за зміну значення другої змінної, хоча обидва значення можуть змінюватися з однаковою швидкістю. Це може статися через невідому третьої змінної, відомої як спотворює фактор. Кореляція передбачає, що обидві змінні незалежні.

З іншого боку, регресія може бути застосована до змінних, які раніше були визначені як залежні і незалежні змінні, і має на увазі, що між цими змінними існує певна ступінь причинності. Причинно-наслідковий зв'язок може бути прямий або непрямої.

У великих даних спочатку може застосовуватися кореляція, для виявлення існування зв'язку. Потім може бути застосована регресія для подальшого вивчення взаємозв'язку і прогнозування значень залежної змінної на основі відомих значень незалежної змінної.

**2. Машинне навчання**

Люди добре розбираються в шаблонах і відносинах всередині даних. На жаль, ми не можемо обробити великі обсяги даних дуже швидко. Машини, з іншого боку, дуже вправні у швидкій обробці великих обсягів даних, але тільки тоді, коли знають як.

Якщо знання людини поєднувати зі швидкістю обробки машини, то останні зможуть обробляти великі обсяги даних, без особливого втручання людини. Це і є основною концепції машинного навчання.

Розглянемо машинне навчання і його взаємозв'язок з Data Mining за допомогою освітлення наступних типів методів машинного навчання:

• Класифікація

• Кластеризація

• Виявлення викидів

• Фільтрація

**Класифікація (машинне навчання з учителем)**

Класифікація - це метод навчання з учителем, при якому дані класифікуються на релевантні, заздалегідь досліджені категорії. Він складається з двох етапів:

1. Система отримує навчальні дані, які вже класифіковані або промарковані, для можливості розвитку розуміння різних категорій.

2. Система отримує невідомі, але аналогічні дані для класифікації, і на підстав розуміння, яке вона отримала з навчальних даних, алгоритм класифікує немарковані дані.

Загальним застосуванням цього методу є фільтрація спаму електронної пошти. Зверніть увагу, що класифікація може проводитися за двома або більше категорій. У спрощеному процесі класифікації машина отримує промарковані дані під час навчання, що дозволяє краще зрозуміти класифікацію. Потім машина отримує немарковані дані, які вона класифікує сама.

Наприклад, банк хоче з'ясувати, який з його клієнтів, найімовірніше, не виконає своїх зобов'язань за кредитом. Ґрунтуючись на історичних даних, складається навчальний набір даних, який містить марковані приклади клієнтів, що раніше виконали або не виконали свої платіжні зобов'язання. Ці навчальні дані передаються в алгоритм класифікації, який використовується для розвитку розуміння «хороших» і «поганих» клієнтів. Нарешті, вводяться нові дані немаркованих клієнтів, щоб з'ясувати, чи належить даний клієнт до категорії, що не виконує своїх зобов'язань.

Приклади питань можуть бути наступними:

• Чи слід прийняти або відхилити заявку претендента на видачу кредитної карти на підстав інших прийнятих або відхилених заявок?

• Чи є томат фруктом або овочем, ґрунтуючись на відомих прикладах фруктів і овочів?

• Результати клінічних випробувань для пацієнтів свідчать про загрозу серцевого нападу?

**Кластеризація (машинне навчання без учителя)**

Кластеризація - це метод навчання без учителя, при якому дані діляться на різні групи, так що дані в кожній групі мають схожі властивості. Не потрібно ніякого попереднього вивчення необхідних категорій. Замість цього категорії генеруються на основі групування даних. Спосіб групування даних залежить від типу використовуваного алгоритму. Кожен алгоритм використовує різну техніку для визначення кластерів.

Кластеризація зазвичай використовується в Data Mining для отримання уявлення про властивості заданого набору даних. Після розробки такого розуміння класифікація може використовуватися для більш точного прогнозування подібних, але нових або попередньо неспостережуваних даних.

Кластеризація може застосовуватися для категоризації невідомих документів і персоналізованих маркетингових кампаній, групуючи клієнтів зі схожим поведінкою.

Наприклад, банк хоче представити своїм існуючим клієнтам ряд нових фінансових продуктів на основі профілів клієнтів, які він має в своїх звітах. Аналітики поділяє клієнтів на кілька груп за допомогою кластеризації. Потім кожній групі надають один або кілька фінансових продуктів, найбільш придатних для характеристик загального профілю групи.

Приклади питань можуть включати в себе:

• Скільки існує різних видів дерев, ґрунтуючись на схожістю між деревами?

• Скільки існує груп клієнтів, ґрунтуючись на історії аналогічних покупок?

• Які різні групи вірусів, ґрунтуючись на їх характеристиках?

**Виявлення викидів**

Виявлення викидів - це процесі пошуку даних, які суттєво різняться від інших даних в заданому наборі даних або несумісні з ними. Цей метод машинного навчання використовується для виявлення аномалій, патологій і відхилень, які можуть бути корисні, наприклад, можливості або невигоди, такі як ризики.

Виявлення викидів тісно пов'язане з концепцією класифікації і кластеризації, хоча алгоритми і зосереджені на пошуку аномальних значень. Процесі може ґрунтуватися або на контрольованому, або на неконтрольованому навчанні. Додатки для виявлення викидів включають виявлення шахрайства, медичну діагностику, аналіз мережевих даних і аналіз даних датчиків.

Наприклад, щоб вияснити, чи є транзакція шахрайською, команда ІТ-фахівців банку будує систему, що використовує метод виявлення викидів, заснована на навчанні з учителем. Спочатку набір відомих шахрайських транзакцій подається в алгоритм виявлення викидів. Потім після навчання системи невідомі транзакції подаються в алгоритм виявлення викиду, щоб передбачити, чи є вони шахрайськими чи ні.

Приклади питань можуть включати в себе наступне:

• Чи є спортсмен тим, хто використовував допінгові препарати?

• Чи є неправильно ідентифіковані фрукти і овочі в навчальному наборі даних, який використовується для завдання класифікації?

• Чи існує особливий штам вірусу, який не реагує на ліки?

**Фільтрація**

Фільтрація - це автоматизований процесі пошуку релевантних елементів з сукупності всіх елементів. Елементи можуть бути відфільтровані або на основі власної поведінки користувача, або шляхом зіставлення поведінки декількох користувачів. Фільтрація зазвичай застосовується за допомогою наступних двох підходів:

• колаборативної фільтрації

• фільтрації на основі контенту

Широко поширеним засобом для здійснення фільтрації, є використання рекомендаційних систем. Колаборативна фільтрація - це метод фільтрації елементів, заснований на взаємодії або злиття попередніх дій користувача з поведінкою інших. Попереднє поведінку досліджуваного користувача, в тому числі і його переваги, рейтинги, історії покупок і багато іншого, взаємодіють з поведінкою подібних користувачів. Ґрунтуючись на схожості поведінки користувачів, елементи фільтруються для досліджуваного користувача.

Колаборативна фільтрація заснована виключно на схожості між поведінкою користувачів. Для точної фільтрації елементів потрібна велика кількість даних про поводжень користувачів. Це є прикладом застосування закону великих чисел.

Фільтрація на основі контенту - це метод фільтрації елементів, сфокусований на схожості між користувачами і елементами. Профіль користувача створюється на основі попередньої поведінки користувача, наприклад, його переваг, оцінок і історії покупок. Подібності, виявлені між профілем користувача і атрибутами різних елементів, призводить до елементів, які фільтруються для користувача. На противагу колаборативній фільтрації, фільтрація на основі контенту призначена виключно для індивідуальних запитів користувача і не вимагає даних про інших користувачів.

Рекомендаційна система передбачає переваги користувача і формує відповідні пропозиції для користувача. Пропозиції зазвичай стосуються рекомендації елементів, таких як фільми, книги, веб-сторінки і люди. Рекомендаційна система зазвичай використовує або колаборативну фільтрацію, або фільтрацію на основі контенту для генерування своїх пропозицій. Вона також може ґрунтуватися на гібриді з колаборативної фільтрації і фільтрації на основі контенту для налагодження точності і ефективності генерованих пропозицій. Наприклад, щоб реалізувати можливості перехресних продажів, банк будує рекомендаційний систему, яка використовує фільтрацію на основі контенту.

На основі запропонованих варіантів між фінансовими продуктами, придбаними клієнтами, і властивостями аналогічних фінансових продуктів рекомендаційна система автоматизує пропозиції для потенційних фінансових продуктів, в яких клієнти також можуть бути зацікавлені.

Приклади питань можуть включати в себе:

• Як можна відображати тільки ті новинні які цікавлять користувача?

• Які місця для відпочинку можна порекомендувати, ґрунтуючись на історії подорожей

• Яких інших нових користувачів можна запропонувати в якості друзів на основі поточного користувача?

**3. Семантичний аналіз**

Фрагмент тексту або мовні дані можуть мати різні значення в різних контекстах, тоді як повне пропозицію може зберегти своє значення, навіть якщо має різну структуру. Для того, щоб машини могли отримувати значиму інформацію, текстові та мовні дані повинні розумітися машинами так само, як і людьми. Семантичний аналіз являє собою практичний метод вилучення значимої інформації з текстових і мовних даних.

Розглянемо наступні типи семантичного аналізу:

• Обробка природної мови

• Обробка тексту

• Аналіз емоційного забарвлення висловлювань

**Обробка природної мови (Natural Language Processing)**

Обробка природної мови - це здатність комп'ютера розуміти людську мову і текст, як це роблять люди. Це дозволяє комп'ютерам виконувати безліч корисних завдань, таких як пошук по всьому тексту.

Наприклад, щоб підвищити якість обслуговування клієнтів, компанія використовує обробку природної мови для перетворення дзвінків клієнтів в текстові дані, які потім аналізуються з метою виявлення причин незадоволеності клієнтів, які часто повторюються.

Замість жорсткого кодування необхідних правил навчання застосовується машинне навчання або з учителем, або без вчителя, для розвитку розуміння комп'ютером природної мови. В цілому, чим більше даних для навчання є в розпорядженні комп'ютера, тим правильніше він зможе розшифрувати текст і мову людини.

Обробка природної мови охоплює як розпізнавання тексту, так і розпізнавання мови. При розпізнаванні мови система намагається осмислити слова, а потім виконати дію, наприклад, цитувати їх в текст.

Приклади питань можуть бути наступними:

• Як може бути розроблена автоматизована система обміну телефонними дзвінками, яка здатна розпізнавати правильні додаткові номери на основі словесного розмови з абонентом?

• Як можна автоматично ідентифікувати граматичні помилки?

• Як спроектувати систему, яка зможе правильно розуміти різні акценти мови?

**Обробка тексту**

Неструктурований текст, як правило, набагато складніше аналізувати і шукати в порівнянні зі структурованим текстом. Обробка тексту - це спеціалізований аналіз тексту за допомогою застосування Data Mining (інтелектуального аналізу даних), машинного навчання та методів обробки природної мови для вилучення цінних знань з неструктурованого тексту. Текстова аналітика забезпечує можливість дослідження тексту, а не просто його пошук.

Корисну інформацію з текстових даних можна отримати, допомагаючи компаніям розвинути розуміння інформації, що міститься в тексті великого об'єму.

Основним принципом обробки тексту є перетворення неструктурованого тексту в дані, які можна шукати і аналізувати. У міру збільшення кількості оцифрованих документів, електронних листів, повідомлень в соціальних мережах і файлів журналів компаніям все частіше доводиться використовувати будь-які цінні відомості, які можна отримати з цих форм слабоструктурованих і неструктурованих даних. Аналіз тільки операційних (структурованих) даних може привести до того, що підприємства упустять можливості економії або розширення бізнесу, особливо ті, що орієнтовані на клієнтів.

Програми включають в себе класифікацію і пошук документів, а також побудова всеосяжного уявлення про клієнта шляхом вилучення інформації з системи CRM.

Обробка тексту зазвичай складається з двох кроків:

1. Аналіз тексту в документах для отримання наступних сутностей:

• Іменовані об'єкти - особистість, група, місце, компанія

• Об'єкти на основі шаблонів - номер соціального страхування, поштовий індекс

• Концепти - абстрактне уявлення об'єкта

• Факти - взаємозв'язок між об'єктами

2. Категоризація документів з використанням цих виокремлених об'єктів і фактів.

Виокремлення інформації може використовуватися для виконання контекстно-залежного пошуку об'єктів, в залежності від типу відносин, який існує між цими об'єктами.

Приклади питань можуть включати в себе наступне:

• Як можна класифікувати веб-сайти, ґрунтуючись на змісті їх веб-сторінок?

• Як можна знайти книги з вмістом, яке має відношення до теми вивчення?

**Аналіз емоційного забарвлення висловлювань**

Аналіз емоційного забарвлення висловлювань - це спеціалізована форма аналізу тексту, яка фокусується на визначенні упередженості або емоцій окремих осіб. Цей вид аналізу визначає ставлення автора тексту, аналізуючи текст в контексті природної мови. Аналіз емоційного забарвлення висловлювань не тільки надає інформацію про те, що відчувають люди, а й інтенсивність їх почуттів. Ця інформація в подальшому може інтегруватися в процесі прийняття рішень. Додатки для аналізу емоційного забарвлення висловлювань охоплюють виявлення задоволеності або незадоволеності клієнтів на ранніх етапах успіху або невдачі продукту або виявлення нових тенденцій.

Прикладами питань можуть бути:

• Як можна виміряти реакцію клієнтів на нову упаковку продукту?

• Який учасник є ймовірним переможцем у пісенному конкурсі?

• Чи може відтік клієнтів визначатися коментарями в соціальних мережах?

**4. Візуальний аналіз**

Візуальний аналіз є формою аналізу даних, яка містить графічне представлення даних для забезпечення або поліпшення їх візуального сприйняття. Ґрунтуючись на припущенні, що людям простіше і швидше зрозуміти і зробити висновки з графіків, ніж з текстів, візуальний аналіз діє як інструмент виявлення знань в області великих даних.

Мета полягає в використанні графічних уявлень для більш глибокого розуміння аналізованих даних. Зокрема, це допомагає виявляв і підкреслювати приховані закономірності, кореляції і аномалії. Візуальний аналіз також безпосередньо пов'язані з пробним аналізом даних, оскільки він підходить різних сторін до формулювання питань.

Розглянемо наступні типи візуального аналізу:

• Кольорові карти

• Тимчасові ряди

• Сітьові графіки

• Зіставлення просторових даних

**Кольорові карти**

Кольорові карти представляють собою ефективний візуальний метод аналізу для вираження шаблонів, композиції даних за допомогою відносин частина-ціле і географічного розподілу даних. Вони також полегшують виявлення кола інтересів і виявлення екстримальних (високих / низьких) значень в наборі даних.

Сама колірна карта являє собою візуальне колірне представлення значень даних. Кожному значенню присвоюється колір відповідно до його типом або діапазоном, в який він потрапляє. Наприклад, колірна карта може привласнювати значення 0-3 червоному кольору, 4-6 - жовтому 7-10 - зеленому.

**Часові ряди**

Графіки часових рядів дозволяють аналізувати дані, які записуються через певні інтервали часу. Цей тип аналізу використовує часові ряди, які представляють собою упорядкований за часом набір значень, записаних через регулярні інтервали часу. Прикладом може бути часовий ряд, що містить дані про продажі, які враховуються в кінці кожного місяця.

Аналіз часових рядів допомагає виявити закономірності в даних, які залежать від часу. Після ідентифікації шаблон можна екстраполюватидля майбутніх прогнозів.

Наприклад, для визначення моделей сезонних продажів, щомісячні показники продажів представляються у вигляді часового ряду, що також допомагає прогнозувати показники продажів в наступному сезоні.

Зазвичай аналіз часових рядів використовується для прогнозування через визначення довгострокових тенденцій, періодичних сезонних моделей і нерегулярних короткочасних відхилень в наборі даних. На відміну від інших видів аналізу аналіз часових рядів завжди включає час в якості порівняльної змінної, і зібрані дані завжди залежать від часу.

Графік часового ряду зазвичай виражається через лінійний графік, з часом, нанесеним на вісь Х, і значенням записаних даних, нанесеним на вісь У.

Приклади питань можуть включати:

• Який урожай слід очікувати фермеру виходячи з історичних даних по врожайності?

• Який очікуваний приріст населення в найближчі 5 років?

• Чи є поточне зниження продажів поодиноким випадком чи воно відбувається регулярно?

**Сітьові графи**

В контексті візуального аналізу сітьовий граф відображає взаємопов'язаний набір об'єктів. Суб'єктом може бути людина, група, або деякий інший об'єкт з бізнес-домену, наприклад продукт. Об'єкти можуть бути пов'язані один з одним безпосередньо або побічно. Деякі поєднання можуть бути тільки односторонніми, так що обхід в зворотному напрямків неможливий.

Сітьовий аналіз - це метод, який фокусується на аналізі відносин між об'єктами в мережі. Він включає в себе побудову об'єктів в вигляді вузлів і зв'язків, таких як ребра між вузлами. Існують спеціалізовані різновиди сітьового аналізу, в тому числі:

• оптимізація маршруту

• аналіз соціальних мереж

• прогнозування поширення, таке як поширення інфекційної хвороби

Приклади питань можуть включати:

• Як можна визначити агента впливу серед великої групи користувачів?

• Чи пов'язані двоє один з одним через довгий ланцюжок родоводу?

• Як я можу визначити закономірності взаємодії між дуже великим числом взаємодій білка?

**Зіставлення просторових даних**

Просторові або геопросторові дані зазвичай використовуються для визначення географічного розташування окремих об'єктів, які потім можуть зіставлятися. Аналіз просторових даних сфокусований на аналізі геолокаційних даних для пошуку різних географічних взаємозв'язків і закономірностей між об'єктами.

Просторові дані обробляються через географічну інформаційну систему, яка відображає просторові дані на карті, в основному, використовуючи координати довготи і широти. Вона надає інструментарій, який дозволяє здійснювати інтерактивне дослідження просторових даних, наприклад, вимірювання відстані між двома точками, або визначення області навколо точки у вигляді кола із заданою відстанню на основі радіусу. Завдяки постійно зростаючої доступності даних на основі визначення їх місця розташування, таких як дані сенсорів і соціальних мереж, просторові дані можуть бити проаналізовані для отримання інформації про місцезнаходження.

Додатки для аналізу просторових даних включають операційну і логістичну оптимізацію, екологічні науки і планування інфраструктури. Дані, що використовуються в якості вхідних для аналізу просторових даних, можуть містити точні місця розташування, такі як довгота і широта, або інформацію, необхідну для розрахунку розташування, наприклад, поштові індекси або ІР-адреси.

Крім того, аналіз просторових даних може використовуватися для визначення кількості об'єктів, які потрапляють в певний радіус іншого об'єкта. Місця розташування визначаються з повідомлень користувачів в соціальних мережах, а персоналізовані пропозиції доставляються в реальному часі в залежності від близькості користувача до об'єкту.

Приклади питань можуть бути наступними:

• Скільки будинків буде знесено через проект розширення дороги?

• На яку відстань повинні переміщатися клієнти, щоб дістатися до об'єкта?

# ТЕМА 5. ВІЗУАЛІЗАЦІЯ ВЕЛИКИХ ДАНИХ

1. Поняття візуалізації. Мова візуалізації

2. Сфера застосування візуалізації та завдання які вона виконує

3. Види візуалізації

4. Коротка характеристика інструментів для візуалізації даних

*Конспект лекції укладено на основі джерел:*

36 кращих інструментів для візуалізації даних. URL: <https://toplead.com.ua/ua/blog/id/38-luchshih-instrumentov-dlja-vizualizacii-dannyh-160/>

Кобзаренко Д.Н., Мустафаев А.Г. Учебное пособие дисциплины«Анализ больших данных» для направления подготовки 38.03.05«Бизнесинформатика», профиль «Электронный бизнес». – Махачкала: ДГУНХ, 2019 г.– 107 с.

Силен Дэвид, Майсман Арно, Али Мохамед Основы Data Science иBig Data. Python и наука о даных. СПб.: Питер, 2017. 336 с.

Як і для чого використовують візуалізацію даних? URL: http://eidos.org.ua/novyny/yak-i-dlya-choho-vykorystovuvaty-vizualizatsiyu-danyh/

**1. Поняття візуалізації. Мова візуалізації**

Візуалізація - це представлення інформації, даних, фактів у візуальній формі.Водночас, візуалізація є мовою, в якій використовуються геометричні об'єкти- точка, лінія, частина поверхні, а також візуальні канали - колір, довжина,орієнтація, розмір. Фактично, мова візуалізації - це продовження звичайноїмови, тому що тексти - її частина.

Одночасно, як і будь яка мова, її базові елементи можна комбінуватибагатьма способами. Проте, не всі комбінації мають сенс. До того ж, різнітипи даних вимагають різних способів їх представлення мовою візуалізації -для них потрібно використовувати різні способи візуального кодування.

З одного боку, може здатися, що це ускладнює задачу інформаційногодизайнера. Насправді, якщо знати мову візуалізації та правила, у який спосібкраще представляти ті чи інші дані, це сильно полегшує роботу - тому щообмежує кількість можливих варіантів.

Ми використаємо найпростішу схему класифікації. За нею, дані поділяютьсяна три типи:

кількісні (quantitative) - все, що можна порахувати та записати учисловій формі;

впорядковані (ordered) - якісні дані, те, що можна розташувати уякомусь порядку - дні тижня, градації шкали оцінювання (наприклад,від "дуже погано" до "дуже добре");

категорійні (categorical) - невпорядковані якісні дані. Практично все,що не відноситься до перших двох типів - назви країн, назви з будьяких наборів, різноманітні типи, тощо.

Елементами мови візуалізації є мітки та візуальні канали.

*Мітки*- це базові графічні елементи (найпростіші геометричні об'єкти):

точка

лінія

площина (на 2D поверхні)

об'ємне тіло (в 3D)

*Канали*- це спосіб, у який ми можемо показати наші позначки. Тобто, миможемо контролювати як буде виглядати позначка, за допомогою такихвізуальних каналів, як:

позиція

розмір

форма

орієнтація

відтінок, насиченість, яскравість (кольору).

Отже, для візуалізації ваших даних, перше, що необхідно зробити

⎯це порахувати кількість змінних (наприклад, скільки колонок є у вашійтаблиці з даними);

⎯визначити для кожної із цих змінних, до якого типу даних вонавідноситься: до кількісних, впорядкованих чи категорійних.Після цього, для кожної змінної ми можемо вибрати мітку та візуальнийканал, який найкраще для неї підійде.

**2. Сфера застосування візуалізації та завдання які вона виконує**

Розглянемо галузі використання візуалізації.

*Статистика та звіти.* Дані за якийсь період часу показуються разом. Наприклад, статичної картинкою в додатку до звіту або налаштованим графіком в сервісі статистики, з можливістю зміни параметрів його відображення.

*Довідкова інформація.* Доповнення до основного тексту, наочно ілюструє його згаданими даними. Наприклад, дати загальне уявлення про динаміку одного з показників, або відобразити якийсь процес і його етапи; може бути - показати структуру якогось явища.

*Інтерактивні сервіси.* Продукти харчування й проекти, в яких інфографіка є частиною функціональності. Так, в якості засобу навігації по сервісах може бути діаграма процесу. Майже все, що пов'язано з роботою з картами в спеціалізованих системах на кшталт диспетчерських і більшої частини комп'ютерних ігор.

*Ілюстрації.* Красиве відображення даних для створення самостійних ілюстрацій.

*Креслення і схеми.* Спеціалізовані документи, що показують структуру і процес роботи складних інженерних та природних систем.

*Експерименти і мистецтво.* Візуалізація даних у вигляді складних і громіздких зображень, які складно «прочитати» побіжно - обсяг даних і взаємозв'язків між ними такий, що потрібно розбиратися з картинкою по частинах; або просто абстрактні зображення, автоматично згенерували. Останнім часом напрямок все більш популярно і періодично виходить за рамки комп'ютерної графіки - наприклад, у вигляді графіків-скульптур.

Виділимо умовно три типи візуалізації.

Наукова візуалізація. При моделюванні різних об'єктів або процесів з'являються великі обсяги даних.

Інформаційна візуалізація. Опис / уявлення якоїсь абстрактної інформації, отриманої при зборі та обробці багаторівневих даних, для аналізу яких необхідно застосовувати різні кількісні та якісні заходи оцінки.

Візуалізація роботи програмного забезпечення.

Візуалізація BigData має певні завдання:

візуалізація потоків даних;

візуальний інтелектуальний аналіз даних;

візуальний пошук і рекомендації;

опис ситуацій на основі великих даних з використанням візуалізації;

масштабовані методи паралельної візуалізації;

сучасні апаратні засоби і архітектури для аналізу і візуалізації даних;

людино-комп'ютерний інтерфейс і візуалізація великих даних;

додатки візуалізації великих даних.

Можна сформулювати вимоги до такого роду візуалізації:

оцінка придатності (адекватності в візуалізації) видів відображення,

природність (звичність для користувачів),

стійкість до масштабування,

можливість виведення надвеликих обсягів даних,

можливості для представлення складних структур, а також об'єктів особливого інтересу, особливих точок, аттракторов, сингулярностей.

**3. Види візуалізації**

Розглянемо традиційні види візуалізації.

Графікі і діаграми

Інфографіка і схеми

Презентація і аналіз даних

Інтерактивний сторітеллінг

Бізнес аналітика і дашборда

Наукова і медична візуалізація

Карти і картограми

**Графіки і діаграми**

Напевно, самий звичний вид візуалізації даних.

Використовується як для презентації даних, так і для аналізу. Зустріти їх можна і на роботі, і в журналі, і в науковому звіті. Зазвичай знання про існуючі типи діаграм і графіків ми отримуємо зі школи або з стандартного набору в Excel. Однак світ графіків і діаграм не обмежується точковим графіком, стовпчиковою і круговою діаграмою. Існують близько 15 загальновідомих типів діаграм, а всього їх понад 60, при цьому їх кількість збільшується з кожним днем люди придумують нові типи для візуалізації складних і незвичайних даних.

**Інфографіка**

Інфографіка стала дуже популярна в останні роки, хоча існують вже давно. Інфографіка відноситься до журналістики даних, де графіки і схеми пояснюють будь-які факти з обраної теми. Зазвичай інфографіка статична і являє собою довге «простирадло» з картинками і текстом. Відмінною особливістю інфографіки є те, що в ній наводяться вже готові висновки, тобто читача проводять за руку з обраної теми і при цьому приправляють це все цифрами і картинками. Часто використовується мальований або анімаційний стиль. Часто використовується не до місця або «для краси», хоча звичайно ж є чудові і цікаві приклади.

**Презентація та аналіз даних**

Один самих звичних способів використання візуалізації даних - презентація інформації у вигляді діаграм або інфографіки. І якщо з цим все зрозуміло, то використання візуалізації для аналізу інформації в основному використовується тільки бізнес-аналітиками та вченими. У чому полягає відмінність?

При аналізі даних за допомогою візуалізації використовують створення великої кількості різних візуальних уявлень одних і тих же даних. Робиться це для можливості знаходження прихованих, на перший погляд, взаємозв'язків і залежностей, а також первинної оцінки набору даних для можливості застосування в подальшому більш складних інструментів аналізу. Цей підхід називається Exploratory data analysis (EDA), що можна перекласти як розвідувальний аналіз даних. Основна відмінність від презентації даних - візуалізація тут може бути «чорновий», але виконується швидко і однією людиною або невеликою робочою групою.

**Інтерактивний сторітеллінг**

Сторітеллінг - це підношення будь-якої корисної інформації в формі цікавої розповіді. Інтерактивний сторітеллінг - розповідь, з яким слухач може взаємодіяти. Користувач може управляти відображенням інформації і знаходити ті залежності, які не знайшов автор. У цьому сенсі він близький до розвідувального аналізу даних, але відрізняється тим, що дані заздалегідь оброблені і представлені в зручному для аналізу вигляді, а також є підказки або заздалегідь прописані сценарії використання.

Тому, найчастіше інтерактивний сторітеллінг називають інтерактивної інфографікою, але для того щоб їй стати недостатньо просто до статичної інфографіку додати спливаючі віконця.

**Дашборди і бізнес аналітика**

Візуалізація активно використовується в бізнесі. Принцип «говорите з даними» допомагає компаніям заробляти більше, а клієнтам отримувати кращий сервіс. Для разового аналізу зазвичай використовується Excel або R. Однак це незручно якщо необхідно стежить за якимись показниками на постійній основі. Для відстеження використовують дашборди - дисплеї, на яких виведені всі необхідні показники в одному місці в вигляді графіків, діаграм і таблиць. Проектування ефективних дашбордів - складна і неординарна завдання. Найчастіше їх перевантажують непотрібною інформацією або намагаються використовувати всі можливі типи шаблонних графіків. Часто для того, щоб спроектувати хороший дашборд, необхідне створення нових типів візуалізації інформації. Тематика активно розвивається за рахунок все більшого застосування аналітики в бізнесі. Також дашборди застосовуються і для особистого використання (фітнес трекери, аналіз особистих витрат і т. П.)

**Візуалізація в медицині та науці**

Специфічний вид візуалізації Його метою зазвичай є виділення закономірностей або аномалій. Від звичайної візуалізації даних відрізняється тим, що часто буває тривимірної і вимагає спеціальної підготовки для інтерпретації.

**Карти і картограми**

Карти - одні з найдавніших способів візуалізації, що відображають навколишню реальність. Картограма - карта з нанесеною на неї інформацією у вигляді кольору або інших способів. Картограми можуть бути використані для відображення будь-якої інформації - від щільності населення, до частоти використання мобільних телефонів в кожному районі країни.

**Хмара тегів**

Кожному елементу в хмарі тегів присвоюється певний ваговий коефіцієнт, який корелює з розміром шрифту. У разі аналізу тексту величина вагового коефіцієнта безпосередньо залежить від частоти вживання (цитування) певного слова або словосполучення.

Дозволяє читачеві в стислі терміни отримати уявлення про ключові моменти скільки завгодно великого тексту або набору текстів.

**Кластерграмма**

Метод візуалізації, що використовується при кластерному аналізі. Показує, як окремі елементи безлічі даних співвідносяться з кластерами в міру зміни їх кількості. Вибір оптимальної кількості кластерів - важлива складова кластерного аналізу.

**Історичний потік**

Допомагає стежити за еволюцією документа, над створенням якого працює одночасно велику кількість авторів. Зокрема, це типова ситуація для сервісів wiki в тому числі. По горизонтальній осі відкладається час. За вертикальної - внесок кожного з співавторів, тобто обсяг введеного тексту. Кожному унікальному автору присвоюється певний колір на діаграмі

**Просторовий потік**

Використовується для відстеження просторового зміни інформації.

**4. Коротка характеристика інструментів для візуалізації даних**

Існує багато спеціальних інструментів для візуалізації: деякі з них зовсім прості: потрібно тільки завантажити дані та вибрати, як вони будуть відображатися. Інші програми більш складні і комплексні — вимагають настройки і, наприклад, знань JavaScript.

[**Plotly**](https://toplead.com.ua/ru/redirect/goto/plot.ly%7C%7C/)

Будує дуже докладні графіки. Ця програма створює діаграми, презентації та дашборди. Ти можеш виконати аналіз за допомогою [JavaScript](https://toplead.com.ua/ru/redirect/goto/uk.wikipedia.org%7C%7Cwiki%7C%7CJavaScript/), [Python](https://toplead.com.ua/ru/redirect/goto/ru.wikipedia.org%7C%7Cwiki%7C%7CPython/), [R](https://toplead.com.ua/ru/redirect/goto/uk.wikipedia.org%7C%7Cwiki%7C%7CR_%28%25D0%25BC%25D0%25BE%25D0%25B2%25D0%25B0_%25D0%25BF%25D1%2580%25D0%25BE%25D0%25B3%25D1%2580%25D0%25B0%25D0%25BC%25D1%2583%25D0%25B2%25D0%25B0%25D0%25BD%25D0%25BD%25D1%258F%29/), [Matlab](https://toplead.com.ua/ru/redirect/goto/ru.wikipedia.org%7C%7Cwiki%7C%7CMATLAB/), Jupyter або Excel. Також є кілька варіантів імпорту даних. Бібліотека візуалізації та інструмент для створення діаграм в режимі онлайн дозволяють створювати по-справжньому красиві графіки.

[**DataHero**](https://toplead.com.ua/ru/redirect/goto/datahero.com%7C%7C/)

Добре підходить, щоб зібрати інформацію з безлічі сервісів в єдину систему. У DataHero можна інтегрувати дані з хмарних сервісів і створювати діаграми та дашборди. Не потребує ніяких спеціальних технічних знань, тому це відмінний інструмент, яким може користуватися вся команда.

[**Chart.js**](https://toplead.com.ua/ru/redirect/goto/www.chartjs.org%7C%7C/)

Чудово підходить для невеликих проєктів. Незважаючи на те, що програма пропонує всього 6 видів діаграм, безкоштовна бібліотека Chart.js підійде для невеликих проєктів. Для побудови діаграм програма використовує [HTML5 Canvas](https://toplead.com.ua/ru/redirect/goto/ru.wikipedia.org%7C%7Cwiki%7C%7CCanvas_%28HTML%29/)і створює швидко реагуючий на зміни простий дизайн.

[**Tableau**](https://toplead.com.ua/ru/redirect/goto/public.tableau.com%7C%7Cs%7C%7C/)

Створює набори даних, якими можна ділитися в режимі реального часу.

Tableau Public — це практично безкоштовний інструмент візуалізації з графіками, діаграмами, картами та іншим. Ви легко зможете завантажити інформацію в систему, а потім спостерігати за тим, як все оновлюється. Для прискорення процесу можна працювати одночасно з іншими учасниками проєкту.

[**Raw**](https://toplead.com.ua/ru/redirect/goto/raw.densitydesign.org%7C%7C/)

Безкоштовний веб-додаток з простим інтерфейсом. Це додаток з відкритим кодом, який можна безкоштовно скачати, змінити і налаштувати під себе. У ньому можна робити векторні візуалізації у форматах SVG або PNG.

[**Dygraphs**](https://toplead.com.ua/ru/redirect/goto/dygraphs.com%7C%7C/)

Підходить для візуалізації великої кількості даних. Це безкоштовний додаток, що дозволяє досліджувати та пояснювати великі обсяги даних. Ви можете налаштувати програму так, як потрібно саме вам, вона працює в усіх основних браузерах. Є функція стиснення графіків для смартфонів і планшетів.

[**ZingChart**](https://toplead.com.ua/ru/redirect/goto/www.zingchart.com%7C%7C/)

Створює діаграми за допомогою HTML5 Canvas. ZingChart — це бібліотека діаграм на JavaScript. Завдяки багатофункціональному API можна створювати інтерактивні Flash або HTML5-діаграми. У програмі понад 100 варіантів діаграм, щоб ви могли вибрати відповідний для ваших цілей і формату даних.

[**InstantAtlas**](https://toplead.com.ua/ru/redirect/goto/www.instantatlas.com%7C%7C/)

Створює гарні візуалізації у вигляді інформаційних карт. Якщо вам потрібен інструмент для візуалізації даних з карт, зверніть увагу на InstantAtlas. У ньому можна створювати інтерактивні динамічні та вузькопрофільні звіти, які об'єднують статистику та картографічну інформацію.

[**Timeline**](https://toplead.com.ua/ru/redirect/goto/www.simile-widgets.org%7C%7Ctimeline%7C%7C/)

Створює інтерактивний таймлайн.Timeline — це зручний віджет, який реагує на рухи мишки користувача. Він спрощує створення графіків з великою кількістю інформації, і видає їх в компактному вигляді. До кожного елементу можна додати більш розгорнуту інформацію, яка буде відображатися при натисканні — жодна деталь не буде упущена.

[**Exhibit**](https://toplead.com.ua/ua/redirect/goto/www.simile-widgets.org%7C%7Cexhibit%7C%7C/)

Перетворює візуалізацію даних на гру. Цей інструмент дозволяє легко створювати інтерактивні карти та інші візуалізації, які можна використовувати в навчальних цілях. Добре підходить для статистичних та історичних наборів даних, таких як прапори різних країн або місця народження відомих людей.

[**Modest Maps**](https://toplead.com.ua/ru/redirect/goto/modestmaps.com%7C%7C/)

У цій програмі можна робити інтерактивні карти та вбудовувати їх на сайт. Цей плагін підходить для дизайнерів, що вважають за краще допрацьовувати функціонал під особисті потреби з урахуванням власного користувальницького досвіду. API підключається досить просто, є можливості для додавання власного коду. Основну бібліотеку можна розширити за допомогою додаткових плагінів із корисними опціями.

[**Leaflet**](https://toplead.com.ua/ru/redirect/goto/leafletjs.com%7C%7C/)

Дозволяє використовувати дані з OpenStreetMap і візуалізувати їх за допомогою HTML5 та CSS3. Ще один інструмент для створення карт, в якому можна створити повністю інтерактивну візуалізацію.

Основна бібліотека сама по собі дуже маленька, але існує величезна кількість плагінів, які розширюють функціонал до рівня профі. Наприклад, можна додати анімовані позначки, маски та зони активності. Ідеально підходить для проєктів, де потрібно показати дані, накладені на географічну розмітку (включаючи нестандартне проєктування).

[**WolframAlpha**](https://toplead.com.ua/ru/redirect/goto/www.wolframalpha.com%7C%7C/)

Дуже добре справляється зі створенням діаграм. Інструмент добре створює діаграми за запитами даних, не потребує додаткового налаштування. Якщо ви хочете візуалізувати загальнодоступні дані, то підійде простий конструктор віджетів.

[**Visual.ly**](https://toplead.com.ua/ru/redirect/goto/visual.ly%7C%7Cproduct%7C%7Cinfographic-design/)

Спрощує візуалізацію даних настільки, наскільки це можливо. Visual.ly — це одночасно і галерея, і інструмент для створення інфографіки. Використовуючи простий набір опцій, можна створювати красиві візуалізації даних. Це не просто візуалізація даних, а щось фантастичне, мрія інфоманіяка!

[**Visualize Free**](https://toplead.com.ua/ru/redirect/goto/www.visualizefree.com%7C%7C/)

Visualize Free — це безкоштовний інструмент, в якому можна використовувати загальнодоступні дані або завантажувати власні і створювати інтерактивні візуалізації. Візуалізації виходять далеко за рамки простих графіків. Для роботи потрібен Flash, але результат може виводитись і в HTML5.

[**FusionCharts**](https://toplead.com.ua/ru/redirect/goto/www.fusioncharts.com%7C%7C/)

Комплексне рішення для побудови діаграм на JavaScript та HTML5. FusionCharts Suite XT пропонує більше 90 графіків і макетів, 965 карт з даними, готові бізнес-панелі та демки. JavaScript API дозволяє легко інтегрувати плагін в будь-який [AJAX](https://toplead.com.ua/ru/redirect/goto/ru.wikipedia.org%7C%7Cwiki%7C%7CAJAX/)-додаток або JavaScript-фреймворк. Діаграми, карти та інформаційні панелі неймовірно інтерактивні, їх легко налаштовувати і вони працюють на всіх пристроях і платформах. У додатку також є порівняльний аналіз топових бібліотек діаграм JavaScript.

[**jqPlot**](https://toplead.com.ua/ru/redirect/goto/www.jqplot.com%7C%7C/)

Чудове рішення для лінійних і точкових діаграм. До плагіну додається кілька приємних додаткових функцій, таких як автоматичне створення трендових ліній та інтерактивних точок, які можуть коригувати відвідувачі сайту, відповідно оновлюючи набір даних.

[**D3.js**](https://toplead.com.ua/ru/redirect/goto/d3js.org%7C%7C/)

Створює незвичайні діаграми. D3.js — це бібліотека JavaScript, що створює діаграми у форматах [HTML](https://toplead.com.ua/ru/redirect/goto/ru.wikipedia.org%7C%7Cwiki%7C%7CHTML/), [SVG](https://toplead.com.ua/ru/redirect/goto/ru.wikipedia.org%7C%7Cwiki%7C%7CSVG/) та [CSS](https://toplead.com.ua/ru/redirect/goto/ru.wikipedia.org%7C%7Cwiki%7C%7CCSS/). Можна використовувати різні джерела даних. Ця бібліотека може сильно підвищити рівень візуалізації складних наборів даних. Програма безкоштовна і використовує веб-стандарти, тому дуже зручна і доступна для користувачів. Також є цікаві варіанти інтерактивної підтримки.

[**JavaScript InfoVis Toolkit**](https://toplead.com.ua/ru/redirect/goto/philogb.github.io%7C%7Cjit%7C%7C/)

Фантастична бібліотека, написана Ніколасом Бельмонте. Модульна структура дозволяє завантажувати тільки те, що абсолютно необхідно для створення візуалізацій. Є ряд унікальних стилів та анімаційних ефектів. Бібліотеку можна використовувати безкоштовно (хоча заохочуються донати).

[**Highcharts**](https://toplead.com.ua/ru/redirect/goto/www.highcharts.com%7C%7C/)

Плагін пропонує великий вибір опцій. Highcharts — це графічна бібліотека JavaScript з величезним діапазоном доступних варіантів діаграм. Результат візуалізується з використанням SVG в сучасних браузерах і VML в Internet Explorer. Графіки автоматично підтримують гарну анімацію, а фреймворк — потоки даних в реальному часі. Highcharts можна завантажити безкоштовно і використовувати в некомерційних цілях (або купити ліцензію для комерційного використання). Також можна відтворювати демки, використовуючи [JSFiddle](https://toplead.com.ua/ru/redirect/goto/ru.wikipedia.org%7C%7Cwiki%7C%7CJsFiddle/).

[**Excel**](https://toplead.com.ua/ua/redirect/goto/products.office.com%7C%7Cen-gb%7C%7Cexcel%3D%3D%3DlegRedir%3Dtrue%26amp%3BCorrelationId%3De00812c4-bc62-4ce2-a9fb-7ab7100a4497/)

Графічно зовсім не гнучкий, але це хороший спосіб вивчити дані. Наприклад, створивши "теплові карти", подібні до цієї. еякі досить складні речі можна робити за допомогою Excel: починаючи з "теплових карт" по клітинам до приблизних діаграм. Як інструмент для початкового рівня він дозволяє швидко вивчити дані або створити візуалізацію для внутрішнього використання. Але є обмеження: стандартний набір кольорів, ліній та стилів ускладнює створення графіки. Проте, він підходить в якості засобу швидкої передачі ідей.

Для цих цілей можна використовувати і електронні таблиці Google. В них можна створювати ті самі діаграми, що і в API Google Chart.

[**Crossfilter**](https://toplead.com.ua/ru/redirect/goto/square.github.io%7C%7Ccrossfilter%7C%7C/)

Кросфільтр в дії: обмежуючи діапазон введення на якомусь одному графіку, ми зачіпаємо всі дані. Це чудовий інструмент для панелей моніторингу або інших інтерактивних інструментів з великими обсягами даних.

У міру того, як з'являються все більш складні інструменти, що дозволяють людям продиратися крізь дані, графіки та діаграми перетворюються на інтерактивні віджети графічного інтерфейсу. Бібліотека JavaScript Crossfilter може бути і першим, і другим. Вона не тільки відображає дані, але і дозволяє побачити реакцію інших пов'язаних діаграм при обмеженні діапазону даних.

[**R**](https://toplead.com.ua/ru/redirect/goto/www.r-project.org%7C%7C/)

Потужна безкоштовна програма для статистичних обчислень і створення графіки. R — це найбільш складний з перерахованих тут інструментів.

Як статистичний збірник, застосовуваний для аналізу великих наборів даних, R — дуже складний інструмент, який вимагає часу на навчання, але пропонує потужну підтримку від інших фахівців та пакетну бібліотеку, яка постійно розширюється. А ще в ньому є власна пошукова система.

Навчитися працювати з цією програмою буде складніше, ніж із будь-якою іншою з перерелічених тут, але це того варте.

[**Weka**](https://toplead.com.ua/ru/redirect/goto/www.cs.waikato.ac.nz%7C%7Cml%7C%7Cweka%7C%7C/)

Weka — це набір алгоритмів машинного навчання для задач інтелектуального аналізу даних. Потужний засіб для вивчення та опрацювання інформації. Weka -— хороший інструмент для класифікації та кластеризації даних, але в ньому можна створювати і прості графіки.

# ТЕМА 6. ВЕЛИКІ ДАНІ ФІНАНСОВОЇ СФЕРИ

1. Місце великих даних у фінансовій сфері

2. Приклади використання великих даних в провідних країнах

3. Використання великих даних в банках України

*Конспект лекції укладено на основі джерел:*

Большие данные в финансовой отрасли: обзор и оценка перспектив развития мирового и российского рынков. URL: [file:///C:/Users/PC/Downloads/idc-26012016.pdf](file:///C%3A/Users/PC/Downloads/idc-26012016.pdf)

Великі дані в індустрії фінансових послуг – від даних до аналізу. URL: <https://www.finextra.com/blogposting/17847/big-data-in-the-financial-services-industry---from-data-to-insights>

Великі дані у фінансах – ваш посібник з аналізу фінансових даних. URL: <https://www.talend.com/resources/big-data-finance/>

Квартальний Н. Великі дані у фінансах – роль аналізу фінансових даних. URL: <https://inoxoft.com/blog/big-data-in-financial-services-role-of-financial-data-analysis-inoxoft/>

Круковець Д. Можливості Data Science в центральних банках: огляд. URL: https://journal.bank.gov.ua/uploads/articles/249\_2\_Krukovets\_Ukr.pdf

Мега Шах. Як використовувати аналітику великих даних для вдосконалення фінансової індустрії URL: https://www.techfunnel.com/fintech/how-to-use-big-data-analytics-to-improve-finance-industry/

Сем Палмер. Аналіз великих даних у фінансових послугах. URL: <https://www.devteam.space/blog/big-data-analytics-in-financial-services/>

Что нужно знать о Big Data в финансах: краткое руководство. URL: <https://aboutdata.ru/2017/06/06/big-data-and-finance/>

Шулин С. Будущее фынансового сектора за Big Data. URL: <https://bluescreen.kz/articles/budushhee-finansovogo-sektora-za-bigdata/>

Що таке великі дані у фінансах. URL: <https://corporatefinanceinstitute.com/resources/knowledge/other/big-data-in-finance/>

Як великі дані покращують банківські та фінансові системи. URL: <https://innovecs.com/blog/big-data-in-banking-and-financial-systems/>

**1. Місце великих даних у фінансовій сфері**

Фінансові послуги взагалі і банківський сектор зокрема належать до вертикальних ринків, на яких стрімке зростання обсягів даних буде і надалі стимулювати впровадження технологій великих даних. За даними досліджень IDC, в Західній Європі рівень поширення технологій великих даних у фінансовій галузі помітно вище середнього і саме в цій галузі плани подальшого впровадження рішень найбільш численні.

Технології великих даних допомагають фінансовим організаціям більш ефективно вирішувати цілий ряд завдань, таких як:

Поліпшення взаємодії з клієнтами. Обмежуючись лише масовим просуванням продуктів, багато банків досі не враховують потреби окремих груп клієнтів. Як наслідок, вони витрачають зайві ресурси, намагаючись продавати не той продукт не тому клієнту. Для вирішення цієї проблеми рекомендується застосування технологій великих даних, які дозволяють ефективно обробляти великі обсяги даних і отримувати корисну інформацію.

Забезпечення відповідності законодавству та галузевим стандартам. Ця проблема не втрачає своєї актуальності для банківського сектора. Вимоги PSD2, Basel III, FACTA - лише деякі приклади нормативи, які продовжують істотно впливати на витрати банків на оптимізацію операційної діяльності. Крім цього, існують і національні нормативи - наприклад, циркуляр Банку Італії №263 зі змінами до вимог інформаційної безпеки і безперервності бізнес-процесів.

Модернізація базових банківських систем. Багато європейських банків потребують ґрунтовному оновленні своїх базових систем, але повністю замінити їх новими системами неможливо через ряд причин. Цей процес буде проходити поетапно, за модульним принципом, починаючи з найбільш застарілих систем.

Впровадження мобільних рішень. Сюди входять мобільний банкінг, мобільні платежі, платежі з використанням технології NFC. Можливість виконання банківських операцій на мобільних пристроях стає однією з найбільш затребуваних послуг, і більшість банків вже займаються розробкою нових інструментів. Очікується, що інвестиції в мобільні послуги триватимуть, але проблеми з доступом до даних, безпекою і цілісністю даних залишаються значними.

Технології великих даних впроваджуються в багатьох галузях - від цільового мобільного маркетингу до виявлення ознак шахрайства та від управління коштами до залучення клієнтів

Фахівці з банківського сектора говорять про те, що віддача від інвестицій в технології великих даних для поліпшення управління персоналом, залучення клієнтів, підвищення операційної ефективності, оптимізації процесів і виявлення ризиків перевершує очікування. До найбільш значущих напрямків використання технологій великих даних відносяться наступні:

* підвищення операційної ефективності;
* поліпшення якості обслуговування клієнтів;
* управління ризиками і дотримання вимог законодавства.

Ось тільки кілька прикладів використання технологій великих даних європейськими фінансовими організаціями: алгоритмічна торгівля (тобто система біржової торгівлі, забезпечує підтримку прийняття рішень про проведення транзакцій на фінансових ринках із застосуванням різноманітних математичних інструментів), аналіз уподобань (аналіз великих обсягів неструктурованих даних, таких як коментарі та дописи в соціальних мережах, з метою оцінки ставлення до тих чи інших брендів, організацій і т.д.) і аналіз чинників впливу (застосування аналітичного інструментарію для передбачення того, який фактор з найбільшою ймовірністю вплине на рішення клієнта).

**2. Приклади використання великих даних в провідних країнах**

Фінансові організації отримують величезні обсяги даних з тисяч різних джерел, і для управління ними зазвичай потрібне спеціальне програмне забезпечення. Звичайний інструментарій, орієнтований на інтелектуальну підтримку бізнесу, призначений для роботи зі структурованими даними, тоді як нове покоління засобів аналітики призначається для роботи з неструктурованими даними з різних джерел. Аналітика, заснована на великих даних, забезпечує оперативний доступ до інформації про бізнес-процесах, подіях і операціях, банк отримує можливість негайно розсилати відповідні повідомлення, оновлювати інформаційні зведення для керівництва, пропонувати стимули готовому покинути його клієнту, перенастроювати банківське обладнання та запобігати шахрайству.

Розглянемо приклад з UniCredit Business Integrated Solutions (UBIS). Через відсутність єдиної точки доступу UBIS доводилося витрачати чимало часу на аналіз даних з багатьох джерел.

Для вирішення цієї проблеми UBIS звернулась до компанії Splunk, постачальника хмарного програмного забезпечення і послуг для пошуку, моніторингу, аналізу та візуалізації великих даних, що надходять від веб-сайтів, додатків, серверів, мереж, датчиків і мобільних пристроїв. Обробка даних, що надходять в режимі реального часу, а також аналіз раніше накопичених обсягів за допомогою рішень Splunk дозволили UBIS моментально виявляти проблеми і проводити профілактику форсмажорних подій. Зокрема, UBIS використовує Splunk для моніторингу транзакцій із заданою періодичністю і видачі сигналів тривоги при досягненні встановлених порогів або настанні певних умов. Завдяки профілактичному моніторингу групі обслуговування клієнтів вдалося значно підвищити якість обслуговування.

Сьогодні бізнес-аналітики UBIS використовують програмне забезпечення Splunk для складання щотижневих звітів для керівництва, що відображають поточну ситуацію за такими виробничими показниками, як:

* кількість клієнтів, що обслуговуються в відділеннях банків з використанням мобільного і інтернет-банкінгу;
* кількість нових відкритих банківських рахунків;
* кількість обслуговуваних позик / кредитних карт;
* кількість проведених платежів / банківських транзакцій.

Для відображення результату операцій мобільного банкінгу на спеціалізованих інформаційних панелях в рішеннях Splunk використовуються карти Google Maps.

Аналіз великих даних відкриває нові можливості підвищення операційної ефективності завдяки доступу в режимі реального часу до інформації. Це дозволяє приймати обґрунтовані рішення і витрачати менше часу на непродуктивні ручні операції. Крім того, застосування засобів великих даних дозволяє перерозподіляти ресурси на користь більш важливих завдань.

Aareal Bank (Німеччина), в 2012 році зайнявся реорганізацією своїх корпоративних даних з метою поліпшення можливості їх аналізу.

Для вирішення цих завдань Aareal Bank звернувся до корпорації SAP. Запропоноване SAP рішення полягало в модернізації вже використовуваного банком додатку SAP NetWeaver Business Warehouse (BW) і перехід на платформу зберігання і обробки даних SAP HANA.

Після реалізації проекту Aareal Bank отримав такі конкретні вигоди:

* отримання більш повного і глибокого уявлення про стан справ в таких галузях, як аналіз продуктів, управління ризиками та фінансами, планування кредитних портфелів і звітність, без звернення за допомогою до відділу інформаційних технологій;
* прискорення підготовки звітів для керівництва на 67%;
* скорочення термінів обробки запитів користувачів в чотири рази;
* підвищення продуктивності праці на 50-70% завдяки скороченню часу очікування відповіді системи, що дозволяє більш оперативно приймати обґрунтовані рішення;
* зниження обсягів сховища даних більш ніж в десять разів

Технології великих даних часто використовуються для ефективного залучення клієнтів. В цьому випадку використовуються дані про результати проведених маркетингових кампаній, дані з систем взаємини з клієнтами (CRM). Аналіз цих даних, а також даних з соціальних мереж, дозволяє виробляти персональний підхід до кожного клієнта або групі клієнтів.

Банк Raiffeisen Bank Austria (RBA) (Хорватія) для підвищення ефективності цільових маркетингових кампаній почав аналізувати клієнтську базу з використанням технологій великих даних. Для цього було вибрано нове CRM-рішення американського постачальника програмних засобів бізнес-аналітики SAS, що включає модель даних банку, а також засоби аналітики для сегментації клієнтської бази, продажу супутніх продуктів, пропозиції альтернативних продуктів і утримання клієнтів. Після впровадження нового CRM-рішення банк зміг побудувати передбачувальні моделі для обслуговування кредитних карт і позик, а потім приступити до запуску кампаній для тестування нових підходів до сегментації клієнтської бази. Коли якість моделей підтвердилося, банк зміг змінити свою маркетингову політику від орієнтації на окремі продукти до орієнтації на клієнта і почати пропонувати клієнтам продукти відповідно до їх потреб і поточними стосунками з банком.

Моніторинг і аналіз даних про банківські транзакції може стати ключовим моментом до підвищення якості обслуговування клієнтів. Одним з ефективних способів отримання інформації про стан програмного забезпечення, що працює на великій кількості різнорідних апаратних платформ, що використовуються клієнтами банку, може виявитися впровадження рішення по збору та аналізу великих даних. В результаті такого аналізу з'являється можливість оперативного виявлення і усунення недоліків, підвищення якості обслуговування клієнтів.

Так італійський банк Sinergia вирішуючи питання найбільш ефективного способу доступу до банкоматів співпрацює з корпорацією NCR і постачальником рішень для великих даних і аналітики Inetco.

Ці компанії запропонували об'єднати рішення для управління банкоматами APTRA Vision виробництва NCR і платформу моніторингу транзакцій і аналітики Insight, розроблену Inetco.

APTRA Vision надає функції моніторингу обладнання, необхідні Sinergia для підтримки в робочому стані зростаючого парку банкоматів різних виробників. INETCO Insight забезпечує відсилання повідомлень в режимі реального часу і повну інформацію по клієнтським транзакціях, що ініціюється через банкомат або будь-яку іншу точку контакту з клієнтом, таку як мобільний додаток, інтернет-кіоск або система електронного банкінгу.

Дане рішення дозволяє зменшити число проблем з виконанням транзакцій, знизити час непрацездатності банкоматів або неможливості їх використання, а також виявляти недоліки в роботі обладнання або програмного забезпечення ще до отримання скарг від клієнтів і визначення підозрілих операцій. Це призводить до підвищення як операційної ефективності, так і рівня задоволеності клієнтів.

Використання технологій великих як інструменту управління ризиками дозволяє відстежувати поведінка клієнтів з метою виявлення підозрілої активності, запобігати шахрайство, а також більш ефективно готувати звіти про діяльність організації для контролюючих органів. Отримання цілісної картини стану справ в організації і забезпечення доступу до архівувати даними дозволяють скоротити витрати на аналіз і прогнозування.

В банку Rabobank Nederland (Нідерланди) постало питання, що до отримання повної картини руху даних і визначенням того, як модифікація однієї з систем позначиться на роботі інших. Ключовим компонентом рішення Rabobank стало створення центрального пункту управління всіма даними Data Portal Financing (DPF). Основний для нього стала платформа інтеграції даних PowerCenter, розроблена американською компанією Informatica. після впровадження цієї платформи все системи стали обмінюватися даними через портал DPF. новий портал забезпечив краще представлення даних і дозволив виконувати набагато більш детальний аналіз, використовуючи більш надійну інформацію, зібрану з різних систем. Аналіз тенденцій на ринку нерухомості і динаміки попиту на нові продукти дозволив Rabobank швидко і точно коригувати свої пропозиції ринку. Дотримання нормативних вимог значно спростилося за рахунок більш наочного відображення потенційних ризиків.

**3. Використання великих даних в банках України**

Банківський сектор в Україні займає провідні позиції у використанні великих даних: банки володіють великим масивом інформації. Ці знання відкривають великі можливості для розвитку, але щоб ними скористатися, потрібен спосіб швидкої обробки, оскільки вручну проаналізувати весь масив практично неможливо.

Важливо зазначити, що використовується статистична інформація й не порушуються закони України «Про Персональні дані» і «Про інформацію». Для того, щоб структурувати, проаналізувати й отримати максимум користі від цих даних, і використовується великі дані.

Найчастіше в банківській сфері система великих даних застосовується для вивчення наявних і потенційних клієнтів банку. Наприклад, банки проводять обробку інформації щодо покупок, соціальних мереж людей, аналізують транзакції. Так з’являється можливість пропонувати актуальний банківський продукт і виключити ненадійних клієнтів. Завдяки всьому цьому банку вдається підвищити свій прибуток.

Сфери застосування великих даних в банківському секторі України.

Основні функції великих даних у банках — це скоринг, запобігання шахрайства та аналіз аудиторії.

1. Аналіз аудиторії. Вивчення цільової аудиторії допомагає створювати популярні пропозиції й оптимізувати їхнє просування. За запитом бізнесу експерти Київстар можуть розробити портрет клієнта, щоб краще пізнати цільовий сегмент, його ключові характеристики та переваги. А Look-alike аудиторія («пошук схожих») допоможе в залученні нових клієнтів і підвищенні ефективності реклами.

2. Скоринг на основі великих даних в банках. Використовуючи скоринг на базі великих даних від Київстар, можна точніше й об’єктивно оцінити потенційних клієнтів на предмет надійності та платоспроможності, навіть якщо в них немає кредитної історії, і виявити ризик шахрайства.

3. Теплові карти й геоаналітика. Ці інструменти допоможуть знайти оптимальні місця для нового відділення, банкомата або термінала, на основі характеристик центрального апарата й місць її скупчення.

4. Таргетована розсилка. Суть таргетованої розсилки в тому, щоб донести інформацію щодо послуг, акцій та товарів тим, кому це цікаво. Для налаштування таргетингу можна вибирати різні сегменти центрального апарату, спираючись на захоплення, географічне положення, вік тощо.

Аналіз аудиторії банку за допомогою великих даних. Що більше бізнес знає про «болі» та бажання своїх наявних і потенційних клієнтів, то з більшою ймовірністю зможе утримати перших і залучити других. Адже серед безлічі пропозицій, представлених на ринку, споживач буде шукати варіант, який максимально відповідає його вимогам. І завдання банку — розробити такий продукт і вчасно запропонувати його зацікавленим людям.

На підставі великих даних від Київстар експерти складають портрет клієнта, щоб краще пізнати потреби й побажання центрального апарату. Для створення портрета Data-фахівці вивчають і аналізують великі дані, визначаючи важливі закономірності. Отримані результати допомагають у розробленні релевантних пропозицій і в запуску ефективнішої реклами для залучення нових клієнтів.

Також великі дані від Київстар дає можливість скористатися інструментом Look-alike — «пошук схожих». Дуже добре підходить для банківської сфери, оскільки ґрунтується на даних з наявної клієнтської бази. Суть процесу в тому, щоб знайти цільову аудиторію, схожу на вже наявних клієнтів. Для цього дата-фахівці, використовуючи методи аналізу Великих даних і машинне навчання, знаходять ключові особливості наявних клієнтів і на їх основі визначають аудиторію, у якій є збіги з клієнтами банку за тими чи іншими критеріями.

Крім того, під час роботи з великими даними можна розрахувати величину життєвого циклу потенційного клієнта (CLV), отже, дізнатися, наскільки вигідним буде співпраця з ним. Так можна прогнозувати свої доходи й постійно їх збільшувати, залучаючи й утримуючи перспективнішу аудиторію.

Скоринг із застосуванням великих даних від Київстар. Щоб зменшити кількість проблемних позичальників банків, дата-фахівці Київстар розробляють [скорингові моделі](https://kyivstar.ua/uk/business/products/antifraud) на основі великих даних. Водночас використовують дані від телеком-оператора, які актуальніші, оскільки на відміну від інформації з Бюро кредитних історій, постійно оновлюються.

Отже, переваги скорингу із застосуванням великих даних від Київстар:

* Створення скорингової моделі на базі актуальної інформації.
* Висока швидкість обробки даних.
* Поліпшення ситуації з видаванням кредитів клієнтам без кредитної історії.
* Надання кредитного та фінансового скорингу.
* Антифрод-скоринг для відсіювання шахраїв.
* Запобігання шахрайства в банках за допомогою великих даних.

Одне з основних напрямів у роботі банків — захист даних і запобігання шахрайства. Для цього використовують антифрод-скоринг, спрямований на виявлення неблагонадійних клієнтів.

Антифрод-скоринг на основі великих даних від Київстар дає можливість охопити більше інформації та зробити прогнози точнішими. Він ефективно працює навіть у таких ризикових фінансових послугах, як мікрокредитування і кредитування онлайн.

Отже, великі дані в банках допомагає за короткі терміни аналізувати й систематизувати великі масиви інформації, познайомитися ближче з цільовою аудиторією, дізнатися, де шукати нових клієнтів і як налагодити з ними комунікацію. Також великі дані від одного з найбільших телеком-операторів України — Київстар — це надійне джерело актуальної інформації для створення ефективних скорингових моделей, розроблення нових продуктів з урахуванням особливостей цільового сегмента й маркетингових кампаній для просування бізнесу.

# ТЕМА 7. БЕЗПЕКА ТА КОНФІДЕНЦІЙНІСТЬ ВЕЛИКИХ ДАНИХ У ФІНАНСАХ

1. Поняття інформаційної безпеки

2. Підходи до безпеки на основі великих даних

3. Безпека та конфіденційність даних у фінансовій сфері (на прикладі банківських установ)

*Конспект лекції укладено на основі джерел:*

Захист персональних даних . URL: <https://www.atcominvestbank.com/uk/about/personal-data-protection>

ІЕ безпека і захист інформації. URL: <https://infotel.ua/ua/IT-bezopasnost-i-zacshita-informatsii-1/>

Ілюк А. Ризики, пов'язані із захистом персональних даних в контексті Big Data. *Юридична газета online*. URL: https://yur-gazeta.com/publications/practice/inshe/riziki-povyazani-iz-zahistom-personalnih-danih-v-konteksti-big-data.html

Константинівна І.О. Сучасні криптосистеми. URL: <https://sites.google.com/site/sucasnikriptosistemik/home>

Основные понятия безопасности для работы с Кластеры больших данных SQL Server URL: <https://docs.microsoft.com/ru-ru/sql/big-data-cluster/concept-security?view=sql-server-ver15>

Охорона банківської таємниці: правові засади. URL: <http://obt.inf.ua/page10.html#top>

Про Систему BankID Національного банку. URL: <https://bank.gov.ua/ua/bank-id-nbu>

Черняк Л. Безопасность больших данных. URL:<https://www.osp.ru/os/2013/02/13034551>

**1. Поняття інформаційної безпеки**

У зв’язку зі зростанням ролі інформаційно-комунікаційних технологій у сучасному суспільстві зростає проблема захисту даних від втрати, викрадення, спотворення або пошкодження, що в свою чергу потребує посиленої уваги. Вирішення цієї проблеми сприяє забезпеченню інформаційної безпеки як окремої особистості, організації, так і всієї держави.

Слово «безпека» латинського походження – secure (securus). Потім в англійській мові воно отримало написання «security».

Інформаційна безпека (information security) – збереження конфіденційності, цілісності та доступності інформації; крім того, можуть враховуватися інші властивості, такі, як автентичність, відстежуваність, неспростовність та надійність.

Інформаційні системиможна розділити на три частини: програмне забезпечення, апаратне забезпечення та комунікації з метою цільового застосування (як механізму захисту і попередження) стандартів інформаційної безпеки. Самі механізми захисту реалізуються на трьох рівнях або шарах: фізичному, особистісному та організаційному. По суті, реалізація політик і процедур безпеки покликана надавати інформацію адміністраторам, користувачам і операторам про те як правильно використовувати готові рішення для підтримки безпеки.

Інформаційна безпека також включає в себе комплекс заходів, які повинні забезпечити захищеність даних від несанкціонованого доступу, використання, оприлюднення, внесення змін чи знищення.

Види загроз інформаційній безпеці:

* отримання доступу до секретних або конфіденційних даних;
* порушення або повне припинення роботи комп’ютерної інформаційної системи;
* отримання доступу до керування роботою комп’ютерної інформаційної системи;
* знищення або спотворення даних.

Останнім часом до питань інформаційної безпеки включено питання інформаційного впливу на особистість і суспільство.

Інформаційна безпека не зводиться винятково до захисту від несанкціонованого доступу до інформації, це ширше поняття, оскільки суб’єкт інформаційних відносин може постраждати (зазнати збитків й/або одержати моральний збиток) не тільки від несанкціонованого доступу, а й від поломки системи, що викликала перерву в роботі.

Спектр інтересів суб’єктів, пов’язаних з використанням інформаційних систем, можна розділити на такі категорії:

* забезпечення доступності інформації;
* забезпечення цілісності інформації;
* забезпечення конфіденційності інформаційних ресурсів і підтримувальної інфраструктури;
* забезпечення вірогідності інформації;
* забезпечення юридичної значимості інформації;
* забезпечення невідстежуваності дій користувача.

**2. Підходи до безпеки на основі великих даних**

До сих пір інформаційна безпека відрізнялася потужними тактичними технологічними засобами на фоні слабкої стратегії, що викликало розквіт кіберзлочинності на тлі відсутності радикальних засобів боротьби з нею - більшість існуючих нині інструментів лікують симптоми, а не причини. Однак є підстави вважати, що ситуація починає змінюватися. Ознаки змін можна побачити насамперед у діях компанії RSA та Cyveillance, які майже одночасно запропонували новий напрямок - ведена (або стимульована) безпека: Intelligence-Driven Security або Intelligence-Led Security.

Такий підхід, радикально відрізняються від традиційних напрямків захисту інформації, до появи яких користувачі займали пасивну позицію по відношенню до загроз і були змушені застосовувати те, що їм пропонували виробники засобів захисту.

Підхід на основі intelligence дозволяє перейти до активної протидії, на основі аналізу відомостей з навколишнього середовища і динамічного реагування на виявлені загрози. Системи безпеки, побудовані на принципах intelligence, дозволять поєднувати оборонні дії з розвідувальними, зі збором і аналізом відомостей із зовнішнього світу. Інакше кажучи, безпека повинна знайти те, що називають системним підходом.

Системне мислення опирається на кілька основних положень:

* будь-яка система складається з окремих частин;
* всі частини системи прямо або побічно пов'язані між собою;
* системи замкнуті, мають свої межі;
* система обмежена в часі і в просторі;
* системи можуть бути вкладені в інші системи або включати в себе інші системи;
* система отримує вхідні впливу ззовні і посилає вихідні впливу зовні, в системі відбуваються процеси, тобто динамічна зміна системи в часі, в тому числі процеси перетворення вхідних чинників у вихідні;
* системи мають властивість синергії.

Поступовий перехід на модель Intelligence-Driven Security дозволяє застосувати технології роботи з Великими Даними з метою інформаційної безпеки. У порівнянні зі звичайними системами SIEM перевага Intelligence-Driven Security складається в можливості аналізувати істотно більший, ніж раніше, обсяг найрізноманітніших, не використаних раніше відомостей. Добре масштабовані системи, побудовані на принципах Intelligence-Driven Security, повинні мати наступні властивості:

* використовувати розвинені підсистеми моніторингу для спостереження за безліччю різноманітних джерел і створення синергетичного ефекту в результаті комбінування відомостей з різних джерел;
* включати автоматизовані засоби для збору і обробки великих даних, підготовки результатів в стандартизованої формі, доступній інших підсистем;
* мати в своєму складі централізоване сховище, потужні аналітичні інструменти і ефективні засоби візуалізації, спільно дозволяють витягти з сировини корисні знання.

У порівнянні зі звичайними системами SIEM перевага Intelligence-Driven Security складається в можливості аналізувати істотно більший, ніж раніше, обсяг найрізноманітніших даних, які раніше залишалися поза увагою через їх начебто незначний вплив на безпеку. Аналітика великих даних дозволяє працювати і з цією групою джерел.

Впровадження нових підходів вплине і на організацію роботи служби інформаційної безпеки, координованої центрами SOC (Security Operations Center). Сьогодні такі центри вже розгорнуті на великих державних і приватних підприємствах. Такі центри називають iSOC (Intelligence-Driven Security Operations Center) - вони повинні бути здатні оцінити всі різноманітні загрози із зовнішнього середовища і пов'язані з ними події для формування цілісного погляду на проблему безпеки.

Передбачається, що їх розвиток буде відбуватися за такими напрямками:

Інтегроване управління інформаційною безпекою. Вже зараз відзначається конвергенція SIEM з системами моніторингу мережі; спільно вони утворюють єдину платформу, здатну аналізувати всю різноманітність даних, що надходять з багатьох джерел. Таким чином вдається об'єднати і уніфікувати потоки вхідних даних, що забезпечує можливість фахівцям-аналітикам працювати з комплексом загроз в цілому, не розпорошуючи свою увагу на окремі прояви.

Управління засобами контролю ідентифікації та прав доступу. Наступне покоління засобів IAM (Identity and Access Management) буде побудовано на постійній оцінці ризиків, що виникають і співвіднесенні цих ризиків з можливостями існуючих засобів захисту. Такі засоби називають націленими на вирішення (solution-aware) - вони здатні оцінити ступінь ризику в поведінці користувача, зокрема при роботі з критично важливими інформаційними ресурсами, навіть після того, як він пройшов процедуру аутентифікації, при цьому враховується все: історія, деталі поточного поведінки і багато іншого, що може дозволити виявити відхилення. При цьому фахівці будуть виходити з того, що заздалегідь неможливо уявити всі сценарії поведінки і можуть бути найнесподіваніші повороти: націлені на вирішення кошти IAM будуть мати більш високий ступінь гнучкості і адаптованості, ніж нинішні жорсткі системи.

Запобігання витоку інформації. Майбутні системи захисту від витоків зможуть розпізнавати фінансові, транзакційні і інші можливі витоки, а їх аналітичні здібності дозволять виявити відмінність між нормальними і шкідливими потоками даних. Основну роль тут відіграватимуть системи, що відносяться до класу «розвідки сесій і аналізу поведінки» (session intelligence and behavior analysis), - наприклад, продукти компанії Silver Tail Systems, нещодавно придбаної RSA, і кошти аналізу поведінки користувачів при роботі з комп'ютером (clickstream analysis ). У зіставленні з бізнес-логікою процесів всі ці кошти дозволять виділити користувачів з аномаліями в поведінці.

Розвиток систем управління ризиками та відповідності нормативним вимогам. У системах GRC (Governance, Risk management, and Compliance) об'єднуються три близькі, але поки що розділені області: загальні принципи керівництва, управління ризиками та підпорядкування регулюючим нормативам. Згодом системи GRC зможуть працювати в режимі реального часу і забезпечать підтримку прийняття керівництвом компаній і менеджментом безпечних рішень.

В результаті аналітика Великих Даних дозволить в повному сенсі цього слова створити кібернетичні системи інформаційної безпеки.

**3. Безпека та конфіденційність даних у фінансовій сфері (на прикладі банківських установ)**

Система управління інформаційною безпекою є сучасним процесом забезпечення безпеки інформаційних ресурсів організації, яка побудована на кращих світових практиках. Стандарти Національного банку України основані на міжнародних стандартах ISO 27001 та ISO 27002 з додаванням вимог із захисту інформації, зумовлених конкретними потребами сфери банківської діяльності і правовими вимогами, які вже висунуто в нормативних документах Національного банку України.

Відповідність системи управління інформаційною безпекою стандартам Національного банку України СОУ Н НБУ 65.1 СУІБ 1.0:2010 та СОУ Н НБУ 65.1 СУІБ 2.0:2010 гарантує банку відповідність міжнародним стандартам ISO 27001 та ISO 27002 і надає можливість отримати відповідний сертифікат.

Необхідність впровадження в банках України стандартів з управління інформаційною безпекою продиктована вимогами Базельського комітету Basel II з управління та зменшення операційних ризиків банків.

Упровадження в банках України стандартів з управління інформаційною безпекою дозволить:

* оптимізувати вартість побудови та підтримання системи інформаційної безпеки;
* постійно відслідковувати та оцінювати ризики з урахуванням цілей бізнесу;
* ефективно виявляти найбільш критичні ризики та знижати ймовірність їх реалізації;
* розробити ефективну політику інформаційної безпеки та забезпечити її якісне виконання;
* ефективно розробляти, впроваджувати та тестувати плани відновлення бізнесу;
* забезпечити розуміння питань інформаційної безпеки керівництвом та всіма працівниками банку;
* забезпечити підвищення репутації та ринкової привабливості банків;
* знизити ризики рейдерських та інших шкідливих для банку атак;
* тощо.

Слід зазначити, що наведені переваги не досягаються шляхом лише «формального» підходу до розроблення, впровадження, функціонування системи управління інформаційною безпекою. Їх досягнення безпосередньо визначається зацікавленістю керівництва і працівників банку в підвищенні рівня інформаційної безпеки.

Крім того впровадження стандартів з питань управління інформаційною безпекою не може бути разовою акцією. Це фактично безперервний процес розроблення, впровадження, функціонування, моніторингу, перегляду, підтримування та вдосконалення СУІБ. Саме тому методологічною основою управління інформаційною безпекою, відповідно до стандартів серії ISO 27000, є процесний підхід.

Для ефективної діяльності організації необхідно ідентифікувати та управляти багатьма видами діяльності. Будь-яку діяльність, що використовує ресурси та підлягає управлінню з метою забезпечення перетворення вхідних даних у вихідні, можна розглядати як процес. Часто вихідні дані одного процесу є безпосередньо вхідними даними для наступного.

Застосування системи процесів у межах організації разом з ідентифікацією цих процесів та їх взаємодіями, а також управління ними можна розглядати як «процесний підхід».

Процесний підхід до управління інформаційною безпекою виводить на перший план важливість:

a) розуміння вимог інформаційної безпеки організації і необхідності розроблення політики та цілей інформаційної безпеки;

б) впровадження заходів безпеки та забезпечення їх функціонування для управління ризиками інформаційної безпеки організації в контексті загальних бізнес-ризиків організації;

в) моніторингу та перегляду продуктивності та ефективності СУІБ та постійного вдосконалення, основаного на об’єктивному вимірюванні.

У межах такого підходу, для процесів СУІБ застосовується модель «Плануй-Виконуй-Перевіряй-Дій» («Plan-Do-Check-Act»), наведена у вступі до стандарту СОУ Н НБУ 65.1 СУІБ 1.0:2010. СУІБ, використовуючи як вхідні дані вимоги інформаційної безпеки та очікування зацікавлених сторін, за допомогою необхідних дій і процесів формує вихідні дані інформаційної безпеки, що відповідають цим вимогам та очікуванням.

Процес управління ризиками інформаційної безпеки повинен здійснюватися для банку в цілому і зокрема включати:

* аналіз і ідентифікацію ризиків;
* оцінку ризиків з точки зору їх впливу на бізнес та ймовірності їх появи;
* інформування особи, яка вправі приймати рішення та акціонерів банку про ймовірності та впливи цих ризиків (ймовірність і наслідки ризику мають бути зрозумілими);
* встановлення порядку та пріоритетів оброблення ризиків;
* становлення пріоритетів виконання дій щодо зниження ризиків;
* участь керівництва в процесі прийняття рішень щодо управління ризиками та його поінформованість щодо стану справ в управлінні ризиками;
* ефективний моніторинг та регулярний перегляд ризиків і процесу управління ризиками;
* інформування керівництва та персоналу щодо ризиків і дій щодо управління ними.

Як приклад забезпечення безпеки і конфіденційності даних наведемо систему BankID Національного банку.

Система BankID Національного банку –це державна система віддаленої ідентифікації, яка забезпечує передачу персональних даних користувачів від банку, в якому відкрито рахунок, до суб'єкта, який надає користувачу послугу.

Використовуючи Систему BankID НБУ громадяни можуть легко та безпечно отримувати доступ до онлайн послуг, які надають:

* державні установи та громадські організації;
* банки та небанківські фінансові установи;
* комерційні установи.

Для користування Системою BankID НБУ громадянам не потрібно додатково реєструватися, достатньо мати відкритий рахунок у банку, що є учасником системи та мати доступ до мобільного або інтернет-банкінгу свого банку.

Електронна дистанційна ідентифікація фізичних осіб з використанням Системи BankID НБУ відбувається шляхом передачі персональних даних такої особи від абонента ідентифікатора (банку, в якому відкрито рахунок користувача), до абонента надавача послуг, який надає послугу користувачу та є безпечною для користувачів. Тільки користувач (власник персональних даних) може ініціювати передачу цих даних від абонента ідентифікатора до абонента надавача послуг. Отже, ніхто, окрім вас, не може ініціювати цей процес! Інформація передається в зашифрованому вигляді відповідно до вимог законодавства України.

Електронна дистанційна ідентифікація фізичних осіб з використанням Системи BankID НБУ відбувається шляхом передачі персональних даних такої особи від абонента – ідентифікатора (банку, в якому відкрито рахунок користувача), до абонента – надавача послуг, який надає послугу користувачу та є безпечною для користувачів.

Тільки користувач (власник персональних даних) може ініціювати передачу цих даних від абонента – ідентифікатора до абонента – надавача послуг. Отже, ніхто, окрім вас, не може ініціювати цей процес! Інформація передається в зашифрованому вигляді відповідно до вимог законодавства України.

* Інтернет-канал передачі даних є захищеним.
* Дані користувачів захищені сертифікатом шифрування та передаються через Систему BankID НБУ із накладанням кваліфікованого електронного підпису або кваліфікованої електронної печатки банку, що передає інформацію.
* Дані передаються виключно одному абоненту – надавачу послуг, де користувач власноруч ініціював запит, тому тільки цей абонент – надавач послуг зможе розшифрувати відповідь з даними від абонента – ідентифікатора.
* Персональні дані користувачів не зберігаються в Системі BankID НБУ.

Звернення фізичної особи до Банку або користування послугами Банку свідчить про згоду такої особи на обробку Банком її персональних даних у зв’язку із таким зверненням чи користуванням послугами Банку.

Банк обробляє персональні дані суб’єктів в тому числі, коли між такими суб’єктами та Банком не укладається окремих письмових правочинів, але вимоги нормативно-правових актів зобов’язують Банк при наданні окремих послуг обробляти персональні дані. Зокрема, у випадках здійснення касових операцій (оплата рахунків, грошові перекази без відкриття рахунків, валютно- обмінні операції тощо) або під час користування послугами банкоматів Банку (отримання готівки з використанням карток, емітентом яких не є Банк).

Обробка персональних даних в такий спосіб може здійснюватись у складі інформаційно-телекомунікаційної системи із застосуванням засобів мережевого захисту від несанкціонованого доступу під час обробки персональних даних.

Склад персональних даних, що обробляється, відповідає складу, що міститься в наданих Банку та підписаних суб’єктом касових платіжних документах або в чеках та квитанціях, які підтверджують здійснення операції.

Додатковими джерелами персональних даних таких суб’єктів можуть бути копії виданих на їх ім'я документів, якщо здійснення копіювання таких документів вимагається нормативно-правовими актами, які є обов’язковими для Банку.

В цих випадках згода суб’єкта на оброблення його персональних даних Банком вважається виявленою ним в зв’язку із здійсненням банківських операцій, отримання банківських послуг, яке супроводжується оформленням касових та/або первинних документів.

Заперечення особи щодо обробки персональних даних, необхідних Банку для виконання своїх зобов’язань, у т.ч. відкликання особою згоди на обробку даних, можуть стати підставою для припинення виконання Банком умов укладених договорів.

У разі відкликання фізичною особою згоди на обробку персональних даних без виконання нею процедур, необхідних для припинення договірних або інших відносин з Банком, Банк продовжуватиме обробку персональних даних в межах та обсягах, обумовлених реалізацією існуючих правовідносин та законодавством України, у т.ч. для захисту Банком своїх прав та законних інтересів за договорами.

# РЕКОМЕНДОВАНА ЛІТЕРАТУРА

**Основна**

1. Chapman, Arthur D. “Principles and methods of data cleaning.” 2005
2. D. Hand, H. Mannila, P. Smyth. 2001. Principles of Data Mining
3. Daniel Keim, Jörn Kohlhammer, Geoffrey Ellis und Florian Mansmann. „Visual Analytics“. 2010
4. Data Science from Scratch by Joel Grus (O‘REILLY) Elements of Statistical Learning: Data Mining, Inference, and Prediction, by Trevor Hastie, Robert Tibshirani and Jerome Friedman
5. Dimitri P. Bertsekas and John N. Tsitsiklis. Introduction to Probability. Charles Wheelan. Naked Statistics: Stripping the Dread from the Data. W. W. Norton and Company, 2013.
6. E. Friedmann, T. Dunning. 2015. Time Series Databases: New Ways to Store and Access Data
7. Guyon, Isabelle, Nada Matic and Vladimir Vapnik. "Discovering Informative Patterns and Data Cleaning." 1996
8. J. Han, M. Kamber. 2011. Data Mining. Concepts and Techniques
9. James Thomas und Kristin Cook. „Illuminating the Path“. 2005 vis.pnnl.gov/pdf/RD\_Agenda\_VisualAnalytics.pdf
10. Osborne, Jason W. “Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data.” 2013
11. Python for Data Analysis by Wes McKinney (O‘REILLY)
12. R. M. Müller, H.-J. Lenz. 2013. Business Intelligence
13. Steven Skiena. “The Data Science Design Manual” <http://www.data-manual.com/>
14. Гурвиц Дж., Ньюджент А., Халпер Ф., Кауфман М. Просто о больших данных: пер. с англ. М.:Эксмо, 2015. 400 с.
15. Кобзаренко Д.Н., Мустафаев А.Г. Учебное пособие дисциплины«Анализ больших данных» для направления подготовки 38.03.05«Бизнесинформатика», профиль «Электронный бизнес». – Махачкала: ДГУНХ, 2019 г.– 107 с.
16. Майер-Шенбергер В. Большие данные. Революция, которая изменит то, как мы живем, работаем имыслим. – М.: Манн, Иванов и Фербер, 2017. – 240 c
17. Радченко И.А, Николаев И.Н. Технологии и инфраструктура Big Data. СПб:Университет ИТМО, 2018. 52 с.
18. Силен Дэвид, Майсман Арно, Али Мохамед Основы Data Science иBig Data. Python и наука о даных. СПб.: Питер, 2017. 336 с.
19. Томас Єрл, Ваджид Хаттак, Пол Булер Основи Big Data: Концепції, алгоритми та технології/Пер.з англ. Анатолія Гладуна;За наук.ред. Олексія Найди. Дніпро: «Баланс Бізнес Букс», 2018. 320 с.
20. Фрэнк Билл. Революция в аналитике. Как в эпоху Big Data улучшить вашбизнес с помощью операционной аналитики М.: Альпина Паблишер, 2014. 430 с
21. Фрэнкс Билл. Укрощение больших данных М.: Манн, Иванов и Фербер, 2014. 352 с.
22. Ын Анналин, Су Кеннет Теоритический минимум по Big Data, что нужно знать о больших даных. СПб.: Питер, 2019. 208 с.

**Допоміжна**

1. 36 кращих інструментів для візуалізації даних. URL: <https://toplead.com.ua/ua/blog/id/38-luchshih-instrumentov-dlja-vizualizacii-dannyh-160/>
2. Big data от А до Я. Часть 3: Приемы и стратегии разработки MapReduce-приложений URL: https://habrahabr.ru/company/dca/blog/270453/ 5
3. BigData от А до Я. Часть 1: Принципы работы с большими данными, парадигма MapReduce URL: <https://habrahabr.ru/company/dca/blog/267361/>
4. Большие данные в финансовой отрасли: обзор и оценка перспектив развития мирового и российского рынков. URL: [file:///C:/Users/PC/Downloads/idc-26012016.pdf](file:///C%3A/Users/PC/Downloads/idc-26012016.pdf)
5. Великі дані в індустрії фінансових послуг – від даних до аналізу. URL: <https://www.finextra.com/blogposting/17847/big-data-in-the-financial-services-industry---from-data-to-insights>
6. Великі дані у фінансах – ваш посібник з аналізу фінансових даних. URL: <https://www.talend.com/resources/big-data-finance/>
7. Вуколов Э.А. Основы статистического анализа. Практикум по статистическим методами исследованию операций с использованием пакетов STATISTICA и EXCEL: учебноепособие для студентов вузов. – М.: ФОРУМ: ИНФРА-М, 2013. – 464 с.
8. Гобарева Я.Л., Городецкая О.Ю., Золотарюк А.В. Бизнес-аналитикасредствами Excel М.: Вузовский учебник, ИНФРА-М, 2013. –336 с.
9. Захист персональних даних . URL: <https://www.atcominvestbank.com/uk/about/personal-data-protection> URL:
10. ІЕ безпека і захист інформації. URL: <https://infotel.ua/ua/IT-bezopasnost-i-zacshita-informatsii-1/>
11. Ілюк А. Ризики, пов'язані із захистом персональних даних в контексті Big Data. *Юридична газета online*. URL: https://yur-gazeta.com/publications/practice/inshe/riziki-povyazani-iz-zahistom-personalnih-danih-v-konteksti-big-data.html
12. Квартальний Н. Великі дані у фінансах – роль аналізу фінансових даних. URL: https://inoxoft.com/blog/big-data-in-financial-services-role-of-financial-data-analysis-inoxoft/
13. Константинівна І.О. Сучасні криптосистеми. URL: <https://sites.google.com/site/sucasnikriptosistemik/home>
14. Круковець Д. Можливості Data Science в центральних банках: огляд. URL: https://journal.bank.gov.ua/uploads/articles/249\_2\_Krukovets\_Ukr.pdf
15. Мега Шах. Як використовувати аналітику великих даних для вдосконалення фінансової індустрії URL: https://www.techfunnel.com/fintech/how-to-use-big-data-analytics-to-improve-finance-industry/
16. Основные понятия безопасности для работы с Кластеры больших данных SQL Server URL: <https://docs.microsoft.com/ru-ru/sql/big-data-cluster/concept-security?view=sql-server-ver15>
17. Охорона банківської таємниці: правові засади. URL: <http://obt.inf.ua/page10.html#top>
18. Про Систему BankID Національного банку. URL: <https://bank.gov.ua/ua/bank-id-nbu>
19. Распределенные базы и хранилища данных : Электронный учебник / Н. Аносова, О. Бородин, Е. Гаврилов и др. – НОУ "ИНТУИТ" URL: <http://www.intuit.ru/studies/courses/1145/214/info>
20. Распределенные файловые системы. Технологии хранения и обработкибольших объемов данных / Computer Science Center URL: https://compscicenter.ru/ courses/ big-data/ 2015-spring/ classes/1117/.
21. Сем Палмер. Аналіз великих даних у фінансових послугах. URL: <https://www.devteam.space/blog/big-data-analytics-in-financial-services/>
22. Ситник В. Ф., Краснюк М. Т. Інтелектуальний аналіз даних (дейтамайнінг): навч. посіб.К.: КНЕУ, 2007. 376 с.
23. Черняк Л. Безопасность больших данных. URL:<https://www.osp.ru/os/2013/02/13034551>
24. Что нужно знать о Big Data в финансах: краткое руководство. URL: <https://aboutdata.ru/2017/06/06/big-data-and-finance/>
25. Шандрівська О. Є., Кириленко А. А.Особливості ідентифікації ризиків ринку big data. URL: http://science.lpnu.ua/sites/default/files/journal-paper/2021/jun/23774/menedzhment121-84-97.pdf
26. Шипунов А.Б., Балдин Е.М., Волкова П.А., Коробейников А.И., Назарова С.А., ПетровС.В., Суфиянов В.Г. Наглядная статистика. Используем R! - М.: ДМК Пресс, 2012. – 298с.
27. Шулин С. Будущее фынансового сектора за Big Data. URL: <https://bluescreen.kz/articles/budushhee-finansovogo-sektora-za-bigdata/>
28. Що таке великі дані у фінансах. URL: <https://corporatefinanceinstitute.com/resources/knowledge/other/big-data-in-finance/>
29. Як великі дані покращують банківські та фінансові системи. URL: <https://innovecs.com/blog/big-data-in-banking-and-financial-systems/>
30. Як і для чого використовують візуалізацію даних? URL: http://eidos.org.ua/novyny/yak-i-dlya-choho-vykorystovuvaty-vizualizatsiyu-danyh/

**Рекомендовані джерела інформації**

1. Законодавчі та нормативно-правові документи. URL: [http://zakon.rada.gov.ua/](http://zakon.rada.gov.ua/7)
2. <https://books.google.com/ngrams>
3. <https://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf>
4. [https://www.data.gov](https://www.data.gov/)
5. <http://data.europa.eu/euodp/de/about>
6. <https://www.govdata.de/>
7. <http://transparenz.hamburg.de/open-data/>
8. <https://dasl.datadescription.com/datafiles/>
9. <http://linkedscience.org/>
10. <https://www.opensciencedatacloud.org/>
11. <https://philogb.github.io/jit/>
12. <http://tagesnetzwerk.de/>
13. [https://tweetping.net](https://tweetping.net/)
14. <https://flowingdata.com/2017/08/18/catalog-of-visualization-types-to-find-the-one-that-fits-your-dataset/>
15. <https://flowingdata.com/2015/12/15/a-day-in-the-life-of-americans/>
16. <https://callingbullshit.org/tools/tools_proportional_ink.html>
17. <http://www.visualisingdata.com/2014/04/the-fine-line-between-confusion-and-deception>
18. <http://tylervigen.com/spurious-correlations>
19. www.domo.com/learn/infographic-data-never-sleeps
20. http://www.originlab.de/
21. http://support.minitab.com
22. https://flowingdata.com/tag/treemap/
23. http://mars.wiwi.hu-berlin.de/mediawiki/teachwiki/index.php/Analysis\_of\_Mortality
24. http://jn.physiology.org/content/104/4/2103
25. http://www.sigchi.org/chi96/proceedings/papers/Tweedie/lt1txt.htm
26. http://rosuda.org/software/Gauguin/gauguin.html