Міністерство освіти і науки України

Державний університет «Житомирська політехніка»

**УПРАВЛІННЯ ВЕЛИКИМИ ДАНИМИ**

**BIG DATA MANAGEMENT**

Методичні вказівки до проведення практичних занять

з дисципліни «Управління великими даними»

для здобувачів вищої освіти «Магістр»

за спеціальністю 051 «Економіка»

освітня програма «Економіка»

Blockchain, Innovation management and Disruptive technology individual assignments are 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

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

УДК 004.6

Управління великими даними [Електронний ресурс] : Методичні вказівки для практичних робіт з курсу «Управління великими даними» для здобувачів вищої освіти за спеціальністю: 051 «Економіка» / уклад. І. І. Сугоняк. – Житомир : Державний університет «Житомирська політехніка», 2021.– Режим доступу:

Методичні вказівки для практичних робіт 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:** Сугоняк Інна Іванівна, кандидат технічних наук, доцент, завідувач кафедри комп’ютерних наук

©Державний університет «Житомирська політехніка», 2021

© І.І. Сугоняк, 2021

**ЗМІСТ**

[ВСТУП 4](#_Toc85712612)

[Практичне заняття 1. ВСТУП В АНАЛІЗ І ОБРОБКУ ВЕЛИКИХ ДАНИХ 5](#_Toc85712613)

[Практичне заняття 2. ОСНОВНІ ПІДХОДИ В РОБОТІ З ВЕЛИКИМИ ДАНИМИ. ЗАГАЛЬНІ МЕТОДИ ОБРОБКИ ВЕЛИКИХ ДАНИХ 7](#_Toc85712614)

[Практичне заняття 3. МЕТОДИ І ТЕХНІКА АНАЛІЗУ BIG DATA. БАЗА ДАНИХ ТА СУБД. ІНСТРУМЕНТИ ОБРОБКИ ВЕЛИКИХ ДАНИХ 9](#_Toc85712615)

[Практичне заняття 4. ДИЗАЙН БАЗИ ДАНИХ. РОЗПОДІЛЕНІ СИСТЕМИ БАЗ ДАНИХ 12](#_Toc85712616)

[Практичне заняття 5. ЗАГАЛЬНІ МЕТОДИ ОБРОБКИ ВЕЛИКИХ ДАНИХ. КОРПОРАТИВНЕ СХОВИЩЕ ДАНИХ У ХМАРІ (WAREHOUSES AND CLOUD) 14](#_Toc85712617)

[Практичне заняття 6. ОСОБЛИВОСТІ ТА АРХІТЕКТУРА HADOOP. ЗБЕРІГАННЯ ТА ОБРОБКА ДАНИХ У HADOOP. КОМПОНЕНТА MAPREDUCE 15](#_Toc85712618)

[ПИТАННЯ ДЛЯ ТЕСТІВ 17](#_Toc85712619)

[СПИСОК РЕКОМЕНДОВАНИХ ДЖЕРЕЛ 19](#_Toc85712620)

# ВСТУП

Управління великими даними – це абсолютно нова концепція управління побудована на постійному розвитку і цифрової трансформації великої кількості даних.

Мета дисципліни «Управління великими даними*»* – познайомити студентів з концепціями управління великими даними, інструментами, що дозволяють виявити дані, підготувати дані, отримати доступ до них, а також механізмами обробки потоків.

Протягом вивчення курсу розглядаються різні техніки і методи аналізу що застосовуються у великих даних, проводиться огляд технологій які збільшили можливості для аналізу великого обсягу даних.

Студентам буде запропоновано роботу з інструментами управління великими даними в економічному середовищі бізнес-процесів, ознайомитися з технологіями обробки великих даних і напрямками їх використання у сучасному світі.

Курс спрямований на формування у студентів компетентностей з питань отримання, обробки, аналізу та оцінювання інформації в різних суспільних сферах на основі управління великими даними та використання отриманих результатів для обґрунтування управлінських рішень.

# Практичне заняття 1. ВСТУП В АНАЛІЗ І ОБРОБКУ ВЕЛИКИХ ДАНИХ

**Питання для обговорення та дискусії**

1. [Що таке великі дані?](https://uk.myservername.com/big-data-tutorial-beginners-what-is-big-data#What_Is_Big_Data)
2. [Проблеми традиційних баз даних](https://uk.myservername.com/big-data-tutorial-beginners-what-is-big-data#Challenges_Of_Traditional_Databases)?
3. [Переваги великих даних перед традиційними базами даних](https://uk.myservername.com/big-data-tutorial-beginners-what-is-big-data#Big_Data_Benefits_Over_Traditional_Database)?
4. [Виклики та ризики в Big Data](https://uk.myservername.com/big-data-tutorial-beginners-what-is-big-data#Challenges_And_Risks_In_BigData)?
5. [Технології великих даних](https://uk.myservername.com/big-data-tutorial-beginners-what-is-big-data#Big_Data_Technologies)?
6. [Інструменти для використання концепцій великих даних](https://uk.myservername.com/big-data-tutorial-beginners-what-is-big-data#Tools_To_Use_Big_Data_Concepts)?
7. [Застосування великих даних](https://uk.myservername.com/big-data-tutorial-beginners-what-is-big-data#Applications_of_Big_Data)?

**Питання для контролю:**

1. Дайте визначення Big Data.
2. Як Big Data дозволяють залучати потенційних клієнтів?
3. В яких галузях застосовують технології Big Data? Наведіть приклади.
4. З яких джерел можуть надходити великі дані?
5. Коли вперше з’явилося поняття Big Data?
6. Чому технології Big Data так важливі для сучасного суспільства?
7. Які завдання практично неможливо вирішити без технологій Big Data?
8. В чому суть ознаки Big Data «об’єм»?
9. В чому суть ознаки Big Data «швидкість»?
10. В чому суть ознаки Big Data «різноманітність»?

**Завдання до виконання**

Початково у сукупність підходів і технологій включались засоби масово-паралельної обробки невизначено-структурованих даних, такі як СУБД NoSQL, алгоритми MapReduce і засоби проекту Hadoop. У подальшому до технологій великих даних почали відносити й інші рішення, що забезпечують схожі за характеристиками можливості обробки надвеликих масивів даних, а також деякі апаратні засоби.

*MapReduce*— модель розподілених обчислювань у комп’ютерних кластерах, представлена компанією Google. Згідно з цією моделлю, додаток розділяється на значну кількість однакових елементарних завдань, що виконуються на вузлах кластера і потім, природнім шляхом зводяться у кінцевий результат.

*NoSQL (від англ. Not Only SQL, не лише SQL)* — загальний термін для різних нереляційних баз даних і сховищ, не означає якусь конкретну технологію чи продукт. Звичайні реляційні бази даних добре підходять для досить швидких і однотипних запитів, а на складних і гнучко побудованих запитах, характерних для великих даних, навантаження перевищує розумні межі і використання СУБД стає неефективним.

*Hadoop*— набор утилітів, бібліотек і фреймворків, що вільно розповсюджується, для розробки і виконання розподілених програм, які працюють на кластерах із сотень і тисяч вузлів. Вважається однією з основоположних технологій більшості даних.

*R*— мова програмування для статистичної обробки даних і роботи з графікою. Широко використовується для аналізу даних і фактично став стандартом для статистичних програм.

*Апаратні рішення.* Корпорації Teradata, EMC та ін. др. пропонують апаратно-програмні комплекси, призначені для обробки великих даних. Ці комплекси поставляються як готові до установки телекомунікаційні шафи, що містять кластер серверів і керівне програмне забезпечення для масово-паралельної обробки. Сюди іноді відносять апаратні рішення для аналітичної обробки в оперативній пам’яті, зокрема, апаратно-програмні комплекси Hana компанії SAP і комплекс Exalytics компанії Oracle, незважаючи на те, що така обробка початково не є масово-паралельною, а об’єми оперативної пам’яті одного вузлаобмежуються кількома терабайтами.

1. Опишіть переваги та недоліки (загрози) технологій Big Data.

2. Створіть хронологічну таблицю розвитку технологій Big Data.

3. Опишіть базовий принцип роботи Big Data – «Горизонтальна масштабованість».

4. Опишіть технології і тенденції роботи з Big Data?

**Підсумковий контроль**

Тестування.

# Практичне заняття 2. ОСНОВНІ ПІДХОДИ В РОБОТІ З ВЕЛИКИМИ ДАНИМИ. ЗАГАЛЬНІ МЕТОДИ ОБРОБКИ ВЕЛИКИХ ДАНИХ

**Питання для обговорення та дискусії**

1. Методи і техніка аналізу великих даних?
2. Поясніть методи класу Data Mining?
3. Що таке змішання та інтеграція даних?
4. Опишіть метод прогнозної аналітики?
5. Опішить метод імітаційного моделювання?
6. Що таке візуалізація аналітичних даних?

**Питання для контролю:**

1. Охарактеризуйте етап життєвого циклу великих даних «ідентифікації даних»
2. Охарактеризуйте етап життєвого циклу великих даних «збір та фільтрація даних»
3. Охарактеризуйте етап життєвого циклу великих даних «перевірка та очищення даних»
4. Охарактеризуйте етап життєвого циклу великих даних «агрегування та подання даних»
5. Які способи очищення та перевірки даних вам відомі?
6. Чи є абсолютно відкритими для аналізу усі дані фірми?
7. Як убезпечитись від недостовірних чи спотворених даних?
8. Для чого проводять агрегування даних?
9. Чи доречним є зберігання даних які відсіяні підчас перевірки для виконання конкретного проекту?

**Завдання до виконання**

Вибір технології MOLAP / ROLAP / HOLAP при аналізі Big Data залежить від частоти оновлення бази даних. З точки зору розпаралелювання обробки, на перший погляд, все просто – будь-який багатовимірний куб може бути «розрізаний» по розподілам одного з вимірів і розподілений між декількома серверами. Наприклад, можна розділити куб на тимчасові періоди (по роках і місяцях), за територіальною ознакою (кожен сервер відповідає за свій регіон) і так далі. Критерієм для поділу куба є наступний принцип: виконання багатовимірного запиту має лягати не на один сервер, а на кілька, після чого отримані результати збираються в єдине ціле. Наприклад, якщо користувач запитує статистику продажів по країні за вказаний проміжок часу, а дані розподілені за кількома регіональним ОЛР-серверів, то кожен сервер повертає свою власну відповідь, які потім збираються воєдино. Якщо ж дані будуть розподілені по тимчасовому критерію, то при виконанні даного прикладу запиту все навантаження ляже на один сервер.

Проблема в тому, що, по-перше, дуже важко заздалегідь визначити оптимальний розподіл даних по серверах, а по-друге, для частини аналітичних запитів може бути заздалегідь невідомо, які дані і з яких серверів знадобляться. Стосовно до Великих Даним це означає, що існуючі підходи для багатовимірного аналізу можуть добре масштабуватися і що вони допускають розподілений збір інформації - кожен сервер може самостійно збирати інформацію, здійснювати її очищення і завантаження в локальну базу.

**Задача 1. Напівструктуровані дані:** Напівструктуровані дані – це дані, які не повністю відформатовані. Він не зберігається в таблицях даних або будь-якій базі даних. Але все-таки ми можемо легко його підготувати та обробити, оскільки ці дані містять теги або значення, розділені комами, тощо. **Приклад** з напівструктурованих даних – це файли XML, файли CSV тощо.

**Неструктуровані дані:** Неструктуровані дані – це дані, які не мають жодної структури. Він може бути у будь-якій формі, не існує заздалегідь визначеної моделі даних. Ми не можемо зберігати його в традиційних базах даних. Її складно шукати та обробляти. Крім того, обсяг неструктурованих даних дуже великий. **Приклад** неструктурованих даних – це тіло електронної пошти, аудіо, відео, зображення, досягнуті документи тощо.

Завдання з відкритих джерел побудувати файл даних або неструктурований або напівструкторований.

**Задача 2.** **Структуровані дані**: Дані, які мають належну структуру або ті, які можна легко зберегти у табличній формі в будь-яких реляційних базах даних, таких як Oracle, SQL Server або MySQL, відомі як структуровані дані. Ми можемо легко та ефективно його обробити або проаналізувати.

Прикладом структурованих даних є дані, що зберігаються в реляційній базі даних, якими можна керувати за допомогою SQL (Мова структурованих запитів). Наприклад, дані працівників (ім'я, ідентифікатор, позначення та заробітна плата) можуть зберігатися у табличному форматі. У традиційній базі даних ми можемо виконувати операції або обробляти неструктуровані або напівструктуровані дані лише після того, як вони відформатовані або вписані в реляційну базу даних. Приклади, структурованих даних є ERP, CRM тощо.

Завдання створити за допомогою комплекс Oracle Exadata створити файл структурованих даних.

**Підсумковий контроль**

Тестування.

# Практичне заняття 3. МЕТОДИ І ТЕХНІКА АНАЛІЗУ BIG DATA. БАЗА ДАНИХ ТА СУБД. ІНСТРУМЕНТИ ОБРОБКИ ВЕЛИКИХ ДАНИХ

**Питання для обговорення та дискусії**

1. Основна складова великих даних
2. Що таке База даних
3. Архітектура Бази даних
4. Класи баз даних
5. Основні інструменти обробки великих даних
6. В чому полягає аналітичний підхід до моделювання?
7. Сутність інформаційного підходу.
8. Назвіть етапи аналізу даних

**Питання для контролю:**

1. Етапи аналізу даних
2. Класифікація моделювання
3. Рух від моделі до результату
4. Механізми зберігання великих даних: їх переваги та недоліки.
5. Бази даних: їх характеристика та перспективи розвитку
6. Плюси та мінуси зберігання даних в оперативній пам’яті
7. Перспективи розвитку та загрози зберігання інформації в хмарних технологіях
8. Назвіть принципи проектування баз даних
9. Які типи СУБД ви знаєте?
10. Охарактеризуйте різні типи СУБД

**Завдання до виконання**

Міжнародна консалтингова компанія McKinsey, що спеціалізується на розв’язанні задач, пов’язаних зі стратегічним управлінням, виділяє 11 методів і технік аналізу, що застосовуються до великих даних.

* Методи классу Data Mining (видобуток даних, інтелектуальний аналіз даних, глибинний аналіз данних) — сукупність методів виявлення у даних раніше невідомих, нетривіальних, практично корисних знань, необхідних для прийняття рішень. До таких методів, зокрема, належать: навчання асоціативним правилам (association rule learning), класифікація (разгалуження на категорії), кластерний аналіз, регресійний аналіз, виявлення і аналіз відхилень тощо.
* Краудсорсинг — класифікація і збагачення даних силами широкого, неозначеного кола особистостей, що виконують цю роботу без вступу у трудові стосунки.
* Змішання та інтеграція даних (data fusion and integration) — набір технік, що дозволяють інтегрувати різнорідні дані з розмаїття джерел з метою проведення глибинного аналізу (наприклад, цифрова обробка сигналів, обробка природньої мови, включно з тональним аналізом).
* Машинне навчання, включаючи навчання з учителем і без учителя – використання моделей, побудованих на базі статистичного аналізучи машинного навчання для отримання комплексних прогнозів на основі базових моделей.
* Штучні нейронні мережі, мережевий аналіз, оптимізація, у тому числі генетичні алгоритми (genetic algorithm — евристичні алгоритми пошуку, що використовуються для розв’язання задач оптимізації і моделювання шляхом випадкового підбору, комбінування і варіації потрібних параметрів з використанням механізмів, аналогічних натуральному відбору у природі)
* Розпізнавання образів
* Прогнозна аналітика
* Імітаційне моделювання (simulation) — метод, що дозволяє будувати моделі, що описують процеси так, як вони би проходили у дійсності. Імітаційне моделювання модна розглядати як різновид експериментальних випробувань.
* Просторовий аналіз (spatial analysis) — клас методів, що використовують топологічну, геометричну і географічну інформацію, що вилучається із даних.
* Статистичний аналіз — аналіз часових рядів, A/B-тестування (A/B testing, split testing — метод маркетингового дослідження; при його використанні контрольна група елементів порівнюється із набором тестових груп, у яких один чи кілька показників були змінені, щоб з’ясувати, які зі змін покращують цільовий показник.
* Візуалізація аналітичних даних — подання інформації у вигляді малюнків, діаграм, з використанням інтерактивних можливостей і анімації, як для отримання результатів, так і для використання у якості вихідних даних для подальшого аналізу. Дуже важливий етап аналізу великих даних, що дозволяє показати найважливіші результати аналізу у найбільш зручному для сприйняття вигляді.

**Задача 1.** **Збір, обробка та структурування даних.** В якості вирішення застосовується Oracle Big Data Appliance – це встановлений Hadoop-кла-стер, Oracle NoSQL Database і засоби інтеграції з іншими сховищами даних. Завдання Oracle Big Data Appliance полягає в зберіганні та первинній обробці неструктурованою або частково структурованої інформації, тобто як раз в тому, що у систем на базі Hadoop виходить найкраще.

Завдання за допомогою хмарного доступу до Oracle Big Data Appliance та відкритих даних здійснити збір та обробку даних за обраним напрямком.

**Задача 2. Агрегація і аналіз даних.**  Для роботи зі структурованими даними використовується комплекс Oracle Exadata. Модулі інтеграції Oracle Big Data Appliance дозволяють оперативно завантажувати дані в Oracle Exadata, а також отримувати доступ до даних «на льоту» з Oracle Exadata.

Завдання використати комплекс Oracle Exadata для завантаження даних.

**Задача 3. Аналітика даних в реальному часі.** Для максимально оперативного аналізу отриманих даних використовується Oracle Exalytics Database Machine, яка дозволяє вирішувати аналітичні завдання. Існує безліч різноманітних методів аналізу масивів даних, в основі яких лежить приблизно однаковий набір інструментів аналізу даних.

Завдання за допомогою ресурсу створити запит і представити вихідні данні. Опишіть методи і техніку аналізу великих даних.

**Підсумковий контроль**

Тестування.

# Практичне заняття 4. ДИЗАЙН БАЗИ ДАНИХ. РОЗПОДІЛЕНІ СИСТЕМИ БАЗ ДАНИХ

**Питання для обговорення та дискусії**

1. Архітектура бази даних: поняття, визначення?
2. Визначення архітектури бази даних?
3. Рівні архітектури Бази даних?
4. Мережеві бази даних?
5. Що таке реляційна модель?

**Питання для контролю:**

1. Що таке об'єктно-орієнтована база даних?
2. Що таке розподілена система баз даних
3. Що таке локальна незалежність?
4. Назвіть аспекти управління розподіленими транзакціями?
5. Принцип апаратної незалежності?
6. Що таке незалежність від операційної системи?
7. Що таке незалежність від мережі?
8. Що таке незалежність від типу БД?

**Завдання до виконання**

З точки зору обробки в основу технологій Big Data покладені два основних принципи:

* розподіленого зберігання даних;
* розподіленої обробки, з урахуванням локальності даних.

Розподілене зберігання вирішує проблему великого обсягу даних, дозволяючи організовувати сховище з довільного числа окремих простих носіїв. Зберігання може бути організовано з різним ступенем надмірності, забезпечуючи стійкість до збоїв окремих носіїв.

Розподілена обробка з урахуванням локальності даних означає, що програма обробки доставляється на обчислювач, що знаходиться якомога ближче до оброблюваних даних. Це принципово відрізняється від традиційного підходу, коли обчислювальні потужності і підсистема зберігання розділені і дані повинні бути доставлені на обчислювач. Таким чином, технології Big Data спираються на обчислювальні кластери з безлічі обчислювачів, забезпечених локальної підсистемою зберігання. Доступ до даних і їх обробка здійснюються спеціальним програмним забезпеченням. Найбільш відомим і інтенсивно розвиваються проектом в області Big Data є Apache Hadoop.

**Задача. Створення архітектури розподіленої бази даних.** Система розподіленої бази даних (DDBMS). Централізована система розподіленого управління базами даних логічно інтегрує дані. Така, що користувач відчуває, що всі дані зберігаються на одному веб-сайті (з ним). DDBMS є єдиним [СУБД](https://uk.itpedia.nl/2017/11/26/wat-is-een-database/) який синхронізує всі дані негайно або періодично, і забезпечити, щоб всі імпорт, оновлення та видалення, виконане виконуються на даних, збережених в іншому місці на сайті автоматично.

Однак централізована база даних має лише один файл бази даних, який знаходиться на одному веб-сайті та доступ до якого здійснюється через єдину мережу.

Особливості розподілених баз даних:

* Якщо колекція розподілених баз даних логічно переплітається, вони фактично утворюють єдину логічну базу даних.
* Сервери на кожному сайті підключені через мережу, але не мають багатопроцесорних функцій.
* Операція незалежна від обладнання.
* Розподілена база даних фізично зберігає дані на декількох сайтах і управляє ними самостійно.
* Операція незалежна від операційної системи.
* Операція незалежна від мережі.
* Нарешті, операція не залежить від СУБД.

Як правило, розподілені бази даних включають такі функції:

* Можливості роботи незалежно.
* Розподілена обробка запитів.
* Розподілена обробка транзакцій.
* Прозора транзакція

Переваги розподіленої бази даних

Використання розподіленої бази даних має багато переваг.

* Якщо в централізованій базі даних виникають збої, система повністю зупиняється. Однак, якщо компонент виходить з ладу в розподіленій системі баз даних, система буде продовжувати працювати зі зниженою продуктивністю до усунення помилки.
* Розподілені системи баз даних дозволяють нам досягти менших витрат на підключення, розміщуючи дані найближче до сайту, який їх найбільше використовує. Це неможливо в централізованій системі.
* Ми можемо розробляти розподілені бази даних модульно. Це означає, що ми можемо розширювати системи, додаючи нові сервери та локальні дані на новий сайт та підключаючи їх до розподіленої системи.

Завдання за допомогою вимог до розподілених баз даних побудувати структуру (архітектуру) бази даних за обраним процесом та детально охарактеризувати її.

**Підсумковий контроль**

Тестування.

# Практичне заняття 5. ЗАГАЛЬНІ МЕТОДИ ОБРОБКИ ВЕЛИКИХ ДАНИХ. КОРПОРАТИВНЕ СХОВИЩЕ ДАНИХ У ХМАРІ (WAREHOUSES AND CLOUD)

**Питання для обговорення та дискусії**

1. Що таке видобуток даних?
2. Що таке Data Mining?
3. Які є етапи аналізу даних?
4. Які види даних можна видобувати?
5. Що таке сховище даних?
6. Які методи використовуються при видобутку даних?

**Питання для контролю:**

1. Яка стратегія чи план тестування для Big Data?
2. Який підхід до тестування застосовується до Big Data?
3. Яке середовище потрібно?
4. Як перевірити Big Data?
5. Які інструменти використовуються при тестуванні Big Data?
6. Що таке OLTP і OLAP?

**Завдання до виконання**

Предметно-орієнтована інформаційна база даних, спеціально розроблена і призначена для підготовки звітів і бізнес-аналізу з метою підтримки прийняття рішень в організації. Будується на базі систем управління базами даних і систем підтримки прийняття рішень. Дані, що надходять в сховище даних, як правило, доступні тільки для читання. Дані з OLTP-системи копіюються в сховище даних таким чином, щоб побудова звітів і OLAP-аналіз не використовував ресурси транзакционной системи і не порушував її стабільність. Як правило, дані завантажуються в сховище з певною періодичністю, тому актуальність даних може трохи відставати від OLTP-системи.

Хмарний сервер (Cloud Server) – це сервер, що надає можливість в будь-який момент збільшити його потужність, якщо в процесі розвитку проект вимагає більше ресурсів. Для збору і обробки Великих даних доцільно використовувати технології хмарних обчислень. Хмарні обчислення – це нова парадигма для розміщення кластерів даних і надання різних послуг локальною мережею або через Інтернет.

Хостинг кластерів даних дає змогу клієнтам зберігати та обчислювати величезну кількість даних у хмарі. Оскільки розмір окремих баз даних зростає швидко і подолав петабайтний бар’єр (наприклад, бази даних соціальних мереж),то онлайн опрацювання в режимі OLAP таких обсягів практично неможливе.

Завдання здійснити порівняльну характеристика OLAP та Big Data.

**Підсумковий контроль**

Тестування.

# Практичне заняття 6. [ОСОБЛИВОСТІ ТА АРХІТЕКТУРА HADOOP](https://uk.myservername.com/what-is-hadoop-apache-hadoop-tutorial#Hadoop_Features). [ЗБЕРІГАННЯ ТА ОБРОБКА ДАНИХ У HADOOP](https://uk.myservername.com/comprehensive-hadoop-testing-tutorial-big-data-testing-guide#Storing_And_Processing_Data_In_Hadoop). КОМПОНЕНТА MAPREDUCE

**Питання для обговорення та дискусії**

1. Що таке MapReduce?
2. Його переваги?
3. Який точний підхід MapReduce?
4. Яка стратегія тестування Big Data?
5. Розробка тестів, виконання тестів Big Data?
6. Управління дефектами та доставка як частина тестування Hadoop?

**Питання для контролю:**

1. Види БД.
2. Які інструменти бізнес-аналітики ви знаєте?
3. В чому полягає сучасний бізнес-аналіз?
4. Яка стратегія або план тестування Big Data?
5. Необхідне середовище для тестування Big Data?
6. Перевірка та перевірка Big Data?
7. Інструменти, що використовуються при тестуванні Big Data?
8. Підхід до тестування Hadoop / Big Data?

**Завдання до виконання**

Ми знаємо, що Hadoop – це те, що використовується для пакетної обробки, і ми також знаємо, що ETL - одне з полів, де Hadoop багато використовується. ETL означає екстракційне перетворення та завантаження . Ми детально обговоримо ці процеси, коли обговоримо план випробувань та стратегію випробувань як точку зору тестування Hadoop.

Відповідно до діаграми, згаданої нижче, ми просто припускаємо, що у нас є чотири різних джерела даних. Операційна система, CRM (Управління відносинами з клієнтами) та ERP (Планування ресурсів підприємства) – це СУБД, або, скажімо, Реляційна система управління базами даних, яка у нас є, і у нас також є кілька плоских файлів, які можуть бути журналами, файлами, записами або чим-небудь стосовно наших джерел даних.

Після того, як дані, що містяться в директорії проміжних етапів, які насправді виявляються HDFS (файлова система розподілу Hadoop). Як тільки Дані будуть трансформовані відповідним чином, використовуючи будь-яку технологію сценаріїв, яка у нас є, ми будемо завантажувати ці дані до сховища даних. З Сховища даних ці дані будуть використовуватися для аналізу OLAP, звітування та видобутку даних або для аналітики.

Які етапи ми можемо використовувати для тестування Hadoop?

Першою фазою буде фаза генерації даних. Тут ми збираємось отримати дані з наших вихідних баз даних або з Flat-файлів, і в такому випадку ми можемо перевірити, що всі Дані були скопійовані успішно та правильно з джерела в директорію проміжного етапу. Він може включати перевірку кількості записів, типу записів і типу полів тощо. Після того, як ці дані скопіюються до директорії проміжних етапів, ми продовжимо і ініціюємо другу фазу, яка є перетворенням (трансформацію). Тут ми матимемо певну бізнес-логіку, яка діятиме на скопійовані дані з вихідних систем і фактично створюватиме або перетворюватиме дані у необхідну бізнес-логіку.

Трансформація може включати сортування даних, фільтрування даних, об’єднання даних із двох різних джерел даних та деякі інші операції. Після того, як Дані перетворені, ми продовжимо і матимемо готові плани випробувань, і ми перевіримо, чи отримуємо вихідні дані, як очікувалось, і всі вихідні дані, які ми отримуємо, відповідають очікуваному результату та типам даних, значенням полів та діапазони тощо – це те, що падає на місце.

Як тільки це буде правильно, ми можемо продовжувати завантажувати дані до сховища даних. На етапі завантаження ми фактично перевіряємо, чи синхронізована кількість записів із етапу та кількість записів у сховищі даних, вони можуть бути не схожими, але вони повинні бути синхронізованими. Ми також бачимо, чи тип даних, які були перетворені, синхронізований.

Повідомляємо, що ми будемо використовувати ці дані для аналізу, звітності та видобутку даних OLAP, що є останнім шаром нашого продукту, і в цьому випадку ми можемо мати наступні або можна сказати, що тестові плани доступні для всіх цих шарів. Але тут, з точки зору тестування Hadoop або Big Data, розробник надасть MapReduce Jobs.

Файли HDFS – будь-які файли, які копіюються в HDFS, інформація про ці файли потрібна для перевірки привілеїв, сценарії HIVE, створені розробниками для перевірки даних у таблиці HIVE, а також нам потрібні UDF HIVE, розроблені розробниками, PIG Сценарії та СВІ СДС. Це все, що нам потрібно отримати від розробників. Перш ніж йти на страту, ми повинні мати усі ці речі.

Для MapReduce Jobs вони нададуть деякі файли JAR, і як частина HDFS вони вже завантажили дані в HDFS, і файли повинні бути готовими, а сценарії HIVE для перевірки даних у таблицях HIVE. Незалежно від того, які СДС вони впровадили, вони будуть доступні в СДІ HIVE. Ми вимагаємо те саме для сценаріїв PIG та UDF.

Завдання визначити який з доступних інструментів роботи з великими даними хмарного типу і що забезпечується Apache може бути використана для функцій аналізу звітності і планування.

**Підсумковий контроль**

Тестування.

# ПИТАННЯ ДЛЯ ТЕСТІВ

1. Дайте визначення Big Data.
2. Як Big Data дозволяють залучати потенційних клієнтів?
3. В яких галузях застосовують технології Big Data? Наведіть приклади.
4. З яких джерел можуть надходити великі дані?
5. Коли вперше з’явилося поняття Big Data?
6. Чому технології Big Data так важливі для сучасного суспільства?
7. Які завдання практично неможливо вирішити без технологій Big Data?
8. В чому суть ознаки Big Data «об’єм»?
9. В чому суть ознаки Big Data «швидкість»?
10. В чому суть ознаки Big Data «різноманітність»?
11. В чому суть ознаки Big Data «достовірність даних»?
12. В чому суть ознаки Big Data «цінність накопиченої інформації»?
13. Що таке принцип «Трьох V»?
14. Дайте визначення структурованих даних.
15. Дайте визначення неструктурованих даних.
16. Дайте визначення напівструктурованих даних.
17. Що таке візуалізація?
18. Для чого використовують візуалізацію?
19. Що є елементами мови візуалізації?
20. Які ви знаєте галузі використання візуалізації?
21. Типи візуалізації.
22. Завдання візуалізації Big Data
23. Вимоги до візуалізації Big Data
24. Традиційні види візуалізації
25. Характеристика інструментів для візуалізації
26. Як використовуються великі дані при пошуку нових клієнтів
27. Порядок застосування великих даних при просуванні нових послуг на ринок
28. Чи можливо негайно перейти до технологій Big Data в усіх сферах фінансового сектору?
29. Що таке цифрова робоча сила у фінансах?
30. Що таке MapReduce?
31. Переваги MapReduce?
32. Який точний підхід MapReduce?
33. [Що таке база даних?](https://uk.itpedia.nl/2017/11/26/wat-is-een-database/)
34. [Що таке SQL?](https://uk.itpedia.nl/2017/11/28/wat-is-sql/)
35. [Що таке NoSQL?](https://uk.itpedia.nl/2017/11/28/wat-is-sql/)
36. Архітектура бази даних: поняття, визначення?
37. Визначення архітектури бази даних?
38. Рівні архітектури Бази даних?
39. Мережеві бази даних?
40. Що таке реляційна модель?
41. Шляхи зниження кредитних ризиків за допомогою технологій Big Data
42. Особливості аналізу великих даних для розширення клієнтської бази
43. Що таке фінансові ризики?
44. Які фінансові послуги використовують великі дані
45. Чи можуть використовувати технології Big Data податкові органи
46. Наскільки аналітика великих даних може підтримувати дотримання фінансових послуг?
47. Підхід до тестування Hadoop / Big Data?
48. Перевірка та перевірка Big Data?
49. Інструменти, що використовуються при тестуванні Big Data?
50. Які інструменти використовуються для аналізу великих даних у фінансовій сфері

**СПИСОК РЕКОМЕНДОВАНИХ ДЖЕРЕЛ**

1. Кузьменко О., Козьменко О. Економіко-математичні методи і моделі. Економетрика : навч. посіб. К. : «Університетська книга», 2019. 406 с.
2. Pang-Ning Tan, Steinbach M., Karpatne A., Kumar V. Introduction to Data Mining. 2nd edition. New York: Pearson, 2018. 1450 p.
3. Zak Cameron. Data Mining Concepts and Techniques: Complete Guide to a Comprehensive Understanding of Data Mining. Independently published, 2020. 366 p.
4. Фрэнкс Б. «Революция в аналитике. Как в эпоху Big Data улучшить ваш бизнес с помощью операционной аналитики» / Билл Фрэнкс. — Москва: Альпина Паблишер, 2017. — 320 c.
5. Глущенко Н. Большие данные большого города: как Big Data меняет жизнь Киева [Электронный ресурс] / Нина Глущенко // интернет-журнал AIN.UA. — Электронные данные. — [Киев: AIN.UA, 2017]. — Режим доступа: <https://ain.ua/special/big-data-in-kyiv/>
6. Силен Д., Мейсман А., Али М. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.
7. Дэви С. Основы Data Science и Big Data. Python и наука о даннях // С. Дэви, М. Арно, А. Мохамед. – СПб.: Питер, 2017. – 336 c.: ил.
8. C. Chen, M. Lin, and X. Guo, “High-level modeling and synthesis of smart sensor networks for Industrial Internet of Things,” Computers & Electrical Engineering, vol. 61, pp. 48–66, 2017.
9. Harrison, Guy. Next Generation Databases: NoSQL, NewSQL and Big Data. 2016.
10. Daniel Keim, Jörn Kohlhammer, Geoffrey Ellis und Florian Mansmann. „Visual Analytics“. 2010
11. 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.
12. F. Liu, Y. Liu, D. Jin, X. Jia, and T. Wang, “Research on Workshop-Based Positioning Technology Based on Internet of Things in Big Data Background,” Complexity, vol. 2018, Article ID 875460, 11 pages, 2018.
13. H. Mora, M. Signes-Pont, D. Gil, and M. Johnsson, “Collaborative Working Architecture for IoT-Based Applications,” Sensors, vol. 18, no. 6, p. 1676, 2018.
14. H. Tahaei, R. Salleh, S. Khan, R. Izard, K.-K. R. Choo, and N. B. Anuar, “A multi-objective software defined network traffic measurement,” Measurement, vol. 95, pp. 317–327, 2017.
15. Hariri, R.H., Fredericks, E.M. & Bowers, K.M. Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data 6, 44 (2019). <https://doi.org/10.1186/s40537-019-0206-3>.
16. [Smart Big Data Management,](https://books.google.de/books?id=A_u3oQEACAAJ&dq=Big+Data+Management&hl=de&newbks=1&newbks_redir=1&sa=X&ved=2ahUKEwiZh-vomq3vAhWMDewKHYk7CgEQ6AEwAnoECAMQAg) Frank Keuper, ‎Dietmar Schmidt, ‎Marc Schomann, 2014.
17. Barrow C., Barrow P., Brown R. The Business Plan Workbook: A Step-By-Step Guide to Creating and Developing a Successful Business. 9th Edition. Kogan Page, 2018. 407 p.
18. Behera H.S. (ed.) et al. Computational Intelligence in Data Mining. Springer, 2018. 895p.
19. Bhatia P. Data Mining and Data Warehousing: Principles and Practical Techniques. Cambridge University Press, 2019. 513 p.
20. Business Result Pre-Intermediate. Teacher’s book. 2nd edition. Rachel Appleby, Mark Bartan, David Grant. Oxford University Press, UK, 2017. 97 p.
21. Byrd M.J. Small Business Management: An Entrepreneur’s Guidebook. 8th Ed. McGraw-Hill, 2017. 494 p.
22. Carlberg C. Business Analysis with Microsoft Excel. 5th edition. Que Publishing, 2018. 576 p.
23. Cicala G. Project Management Using Microsoft Project 2019: A Training and Reference Guide for Project Managers Using Standard, Professional, Server, Web Application and Project Online for Office 365. Project Assistants Inc., 2019. 446 p.
24. Dr. Eckroth J. AI Blueprints: How to build and deploy AI business projects. Packt, 2018. 241p.
25. Esposito A., Esposito A.M., Jain L.C. (Eds.) Innovations in Big Data Mining and Embedded Knowledge. Springer, 2019. 286 p. (Intelligent Systems Reference Library 159).
26. Fitzgerald Brian, Stol Klaas-Jan. Scaling a Software Business. Springer, 2017. 264 p.
27. Harris Tom. Start-up: A Practical Guide to Starting and Running a New Business. Springer, 2018. 153 p.
28. Harvard Business Review – Entrepreneur’s Handbook. Boston: Harvard Business School Publishing Corporation, 2018. 303 p.
29. Hastie Trevor, Tibshirani Robert, Friedman Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009. Corrected 12th printing, 2017. 745 p.
30. Graph Databases, 2nd Edition, by Ian Robinson, Jim Webber, Emil Eifrem, 2015, O'Reilly Media
31. Kreinovich V., Sriboonchitta S., Chakpitak N. (eds.) Predictive Econometrics and Big Data. Springer, 2018. 788 p.
32. Kryvinska N., Gregus M. (eds.) Data-Centric Business and Applications Evolvements in Business Information Processing and Management (Vol.2). Springer, 2020. 468 p.
33. Lee C.F., Chen H.Y., Lee J. Financial Econometrics, Mathematics and Statistics: Theory, Method and Application. Springer, 2019. 656 p.
34. Olson D.L., Wu D. Predictive Data Mining Models. 2nd.ed. Springer, 2020. 125 p. Computational Risk Management.
35. Olson David L. Descriptive Data Mining. 2nd.ed. Springer, 2019. 130 p. (Computational Risk Management).
36. Project Management Institute. A Guide to the Project Management Body of Knowledge: PMBOK Guide-Sixth Edition + Agile Practice Guide. 6th Edition. Project Management Institute, 2017. 800 p.
37. Project Management Institute. Agile: Practice Guide. Project Management Institute, 2017. 115 p.
38. Rojas I., Pomares H., Valenzuela O. (eds.) Advances in Time Series Analysis and Forecasting: Selected Contributions from ITISE 2016. Springer, 2017. 412 p.
39. Sul D. Panel Data Econometrics: Common Factor Analysis for Empirical Researchers. New York: Routledge, 2019. 165 p.
40. Wilcox J. Excel: The Ultimate Statistics Guide. Mobo Publications, 2017. 251p.

***Список інтернет джерел:***

1. <https://rb.ru/howto/chto-takoe-big-data/>
2. <https://www.mysql.com/>
3. <https://app.diagrams.net/>
4. <https://cloudstack.apache.org/>
5. <https://projects.apache.org/projects.html?category>
6. <https://cloud.google.com/bigtable>
7. <https://aws.amazon.com/de/dynamodb/>
8. <https://neo4j.com/>
9. https://www.datacenterknowledge.com/archives/2015/03/30/big-data-bubble-set-burst
10. <https://ru.wikipedia.org/wiki/Большие_данные>
11. <https://intellect.ml/big-data-6821>
12. <http://sewiki.ru/index.php?title=Большие_данные&oldid=3075>
13. ttp://www.mckinsey.com/insights/business\_technology/big\_data\_the\_next\_fro ntier\_for\_innovation
14. <http://engjournal.ru/articles/1228/1228.pdf>
15. http://www.ogcs.com.ua/index.php/articles/121-big-data-v-promyshlennosti-innovatsii-k-kotorym-pridetsya-privykat