НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

Кафедра інформаційних систем і технологій

"**ЗАТВЕРДЖУЮ**"
Декан факультету інформаційних технологій

_____ ____Глазунова О.Г.
"____"____ 2020 р.

РОЗГЛЯНУТО І СХВАЛЕНО
на засіданні кафедри інформаційних систем і технологій
Протокол № 4 від "22" 04 2020 р.
Завідувач кафедри
_____ Швиденко М.З.

РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ «АНАЛІТИКА ВЕЛИКИХ ДАНИХ»

галузь знань 05 «Соціальні та поведінкові науки»
спеціальність 051 «Економіка»
освітня програма «Економічна кібернетика»
факультет Інформаційних технологій
розробники: доц., к.е.н. Харченко В.В.

Київ – 2020 р.

1. Опис навчальної дисципліни «Аналітика великих даних»

Галузь знань, напрям підготовки, спеціальність, освітньо-кваліфікаційний рівень		
Освітній рівень	*Магістр*	
Галузь знань	*05 «Соціальні та поведінкові науки»*	
Спеціальність	*051 «Економіка»*	
Освітня програма	*«Економічна кібернетика»*	
Характеристика навчальної дисципліни		
Вид	*Вибіркова*	
Загальна кількість годин	*135 год.*	
Кількість кредитів ECTS	*4 ECTS*	
Кількість змістових модулів	*2*	
Курсовий проект (робота) (за наявності)	–	
Форма контролю	*Іспит*	
Показники навчальної дисципліни для денної та заочної форм навчання		
	денна форма навчання	заочна форма навчання
Рік підготовки (курс)	*1*	-
Семестр	*2*	-
Лекційні заняття	*30 год.*	-
Практичні, семінарські заняття	-	-
Лабораторні заняття	*30 год.*	-
Самостійна робота	*75 год.*	-
Індивідуальні завдання	-	-
Кількість тижневих аудиторних годин для денної форми навчання	*12 год.*	

2. Мета та завдання навчальної дисципліни

Мета сформувати у студентів фундаментальні знання з теорії та практики в області розробки і використання систем обробки та аналізу великих масивів даних.

Завдання вивчення теоретико-методичних засад та основних технологій щодо вирішення завдань обробки великих за обсягом, швидко змінюваних та погано структурованих даних, що об'єднуються терміном «великі дані».

У результаті вивчення навчальної дисципліни студент повинен

знати: основні поняття аналітики великих даних; основні технології, що застосовуються для зберігання і пошуку в великих даних.

вміти: застосовувати методи аналізу великих даних, вміти реалізовувати програми для аналітики великих даних.

Навчальна дисципліна забезпечує формування ряду фахових компетентностей:

СК3. Здатність збирати, аналізувати та обробляти статистичні дані, науково-аналітичні матеріали, які необхідні для розв'язання комплексних економічних проблем, робити на їх основі обґрунтовані висновки.

СК4. Здатність використовувати сучасні інформаційні технології, методи та прийоми дослідження економічних та соціальних процесів, адекватні встановленим потребам дослідження.

СК11. Здатність створювати та оцінювати моделі економічних процесів як аналітично так і з використанням універсальних програмних засобів і аналітичних платформ, що застосовуються для аналізу даних.

У результаті вивчення навчальної дисципліни студент набуде певні програмні результати, а саме:

ПР 17. Застосовувати сучасні інформаційні системи на підприємствах (установах) різних сфер діяльності, зокрема в аграрній сфері.

3. Програма та структура навчальної дисципліни для повного терміну денної форми навчання:

Змістовий модуль 1. Введення в аналітику великих даних

Тема 1 Введення в аналітику великих даних
Основні поняття та визначення. Історія розвитку. Джерела великих даних. Застосування в економіці, бізнесі, сільському господарстві, промисловості. Приклади використання. Великі дані в наукових сферах. Особливості застосування. Вимоги до професії аналітика великих даних.

Тема 2 Життєвий цикл проекту по аналітиці великих даних
Основні етапи життєвого циклу. Збір, консолідація і очищення даних. Побудова моделей, роль машинного навчання.

Тема 3 Основні техніки (підходи) щодо роботи з великими даними

Збір та консолідація даних, «аналітична пісочниця» (analytic sandbox) «озеро даних» (data lake), пакетна аналітика (batch oriented), аналітика реального часу (real time oriented), гібридна аналітика (hybrid), робота з СУБД.

Тема 4 Введення в когнітивний аналіз даних

Когнітивна система типу IBM Watson. Функції та можливості системи IBM Watson.

Змістовий модуль 2. Технології та інструменти роботи з великими даними

Тема 5 Аналіз та візуалізація великих даних

Візуалізація великих даних («big data visualization»), візуалізація текстів, візуалізація кластерів, візуалізація асоціацій, ландшафтна візуалізація, візуалізація гіпотез, візуалізація дерев рішень, багатовимірна візуалізація. Сіткова візуалізація. Класифікація. Gephi.

Тема 6 Основні технології та інструменти роботи з великими даними

Підхід Map/Reduce та його програмна реалізація, Apache Hadoop, HDFS, HBase, YARN, Hive, Pig, Storm як система потокової обробки, мова програмування Python, R, Apache Spark. IBM Bluemix. Microsoft HDInsight.

Тема 7 No SQL БД та візуалізація великих даних

Реляційні та No SQL БД: характеристика та відмінності. Оцінка «великих даних»: проблеми та вирішення.

Тема 8 Apache Spark та перспективи розвитку великих даних

Принципи роботи технології Apache Spark. Перспективи розвитку великих даних.

Структура навчальної дисципліни

Назви змістових модулів і тем	Кількість годин												
	денна форма							Заочна форма					
	тижні	усього	у тому числі					усього	у тому числі				
			л	п	лаб	інд	с.р.		л	п	лаб	інд	с.р.
1	2	3	4	5	6	7	8	9	10	11	12	13	14
Змістовий модуль 1. Введення в аналітику великих даних													
Тема 1. Введення в аналітику великих даних	1,2	16	2		4		10						
Тема 2. Життєвий цикл проекту аналітики великих даних	3,4	18	4		4		10						
Тема 3. Основні техніки (підходи) аналітики великих даних	5,6	18	4		4		10						
Тема 4. Види аналітики великих даних	7,8	18	4		4		10						
Разом за змістовим модулем 1	70		14		16		40						
Змістовий модуль 2. Технології та інструменти роботи з великими даними													
Тема 5 Основні технології та інструменти роботи з великими даними	9,10	18	4		4		10						
Тема 6 Модель обчислень Map Reduce, Pig, Hive	11,13	18	4		4		10						
Тема 7 No SQL БД та візуалізація великих даних	14	18	4		4		10						
Тема 8 Apache Spark та перспективи розвитку великих даних	15	11	4		2		5						
Разом за змістовим модулем 2	65		16		14		35						
Усього годин	**135**		**30**		**30**		**75**						
Курсовий проект (робота)	**-**		**-**	**-**	**-**		**-**		**-**	**-**	**-**		**-**
Усього годин	**135**		**30**		**30**		**75**						

4. Теми семінарських занять

Не передбачені навчальним планом.

5. Теми практичних занять

Не передбачені навчальним планом.

6. Теми лабораторних занять

№ з/п	Назва теми	Кількість годин
1	Готові рішення аналізу даних (Rapid Miner, Weka), мови Python та R, стек бібліотек аналізу даних.	4
2	Візуалізація даних. Gephi.	4
3	Хмарний сервіс Big Data IBM Bluemix, Azure HD Insights.	4
4	Налаштування кластеру Hadoop for Analytics.	6
5	Файлова система HDFS, Object Storage.	4
6	Модель Map/Reduce, Apache Spark.	4
7	NewSQL платформа SAP HANA, Oracle Exalytics.	4
8	Всього	30

7. Контрольні питання, комплекти тестів для визначення рівня засвоєння знань студентами

1. Основні характеристики великих даних
2. Роль великих даних в сільському господарстві
3. Консолідація даних
4. Візуалізація даних, Gaphi.
5. Основні конструкції мови R, консолідація даних, візуалізація
6. HDFS – основи організації
7. Архітектура Hadoop
8. Виконання Map/Reduce
9. Виконання програм в Hadoop
10. Основи YARN
11. Аналітика потокових даних в платформі Storm
12. Архітектура Apache Spark
13. Організація даних в Apache Spark
14. Обробка даних в GraphX
15. Алгоритми класифікації
16. Алгоритми кластеризації
17. Нейронні мережі як реалізація алгоритмів машинного навчання
18. Інтелектуальне алгоритми
19. Застосування технологій великих даних для задач управління в реальному часі

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

ОС Магістр **Освітня програма** “Економічна кібернетика”	**Кафедра інформаційних систем і технологій** 20__- 20__навч. рік	**ЕКЗАМЕНАЦІЙНИЙ БІЛЕТ № 1** з дисципліни “ **Аналітика великих даних** ”	**Затверджую** Зав. кафедри _____ Швиденко Михайло Зіновійович ___.____.____р.

Екзаменаційні запитання

1. Опишіть принцип роботи MapReduce.

2. HDFS – основи організації.

10 тестових завдань

1. Виберіть чотири головні характеристики Big Data:
1 Virtualization, Volume, Variability, Vehicle
2 Variety, Velocity, Volume, Value
3 Verification, Volume, Velocity, Visualization
4 Video, Value, Variety, Volume

2. Виберіть важливі події, що вплинули на формування тренда великих даних:
1 Розробка мови Pyhton
2 Розробка фреймворку Hadoop
3 Винахід принципу MapReduce
4 Розробка мови Java
5 Розробка мови R

3. Виберіть варіант, в яких дані структуровані
1 Дані щодо продажів підприємства, які наведені у вигляді помісячних звітів у форматі MS Word
2 Таблиця з щоденними показами температури приміщення за рік в файлі формату csv
3 Текст наукового посібника, що наведений у форматі PDF
4 Бібліотека фільмів, представлених у форматі mpeg4 на одному жорсткому диску

4. Розставте послідовність етапів проекту аналітики відповідно до методології CRISP-DM

1)	а) Розуміння бізнес-цілей (Business understanding);
2)	б) Впровадження (Deployment);
3)	в) Моделювання (Modeling);
4)	г) Початкове вивчення даних (Data Understanding);
5)	д) Оцінка (Evaluation)
6)	е) Підготовка даних (Data Preparation)

5. Hadoop – це:
1. Розподілена СУБД, що дозволяє обробляти великі дані
2. Набір утиліт, а також програмний каркас для виконання розподілених програм, що працюють на кластерах
3. Мова виконання завдань відповідно до парадигми MapReduce
4. Розподілена файлова система, що призначена для зберігання файлів великого обсягу

6. Для чого аналітику необхідна «аналітична пісочниця»?
1. Для високопродуктивної аналітики за рахунок використання оперативної пам'яті та inDB операцій
2. Для зберігання всіх отриманих від замовника даних
3. Для побудови звітів про результати аналізу
4. Для зниження витрат, пов'язаних з реплікацією даних

7. Які з перелічених засобів доцільно використовувати для аналізу даних, що представлені єдиним csv-файлом розміром більшим за 100 ГБ
1. Data Warehouse
2. Hadoop
3. «Аналітична пісочниця»
4. Python
5. PHP

8. На якому з етапів процесу CRISP-DM відбувається перевірка якості даних?
1. Розуміння бізнес-цілей (Business understanding)
2. Початкове вивчення даних (Data Understanding)
3. Підготовка даних (Data Preparation)
4. Моделювання (Modeling)
5. Оцінка (Evaluation)
6. Впровадження (Deployment)

9. До слабких сторін Hadoop можна віднести. Виберіть зайве.
1. Hadoop це фреймворк, а не готове рішення
2. Hive та Pig мають безліч архітектурних обмежень порівняно з реляційною СУБД
3. Hadoop легко встановити і налаштувати, але складніше супроводжувати;
4. Hadoop обробляє тільки дуже велику кількість даних, які зберігаються розподілено на безлічі вузлів
5. У деяких випадках Hadoop дуже повільний

10. Принцип MapReduce полягає в тому, щоб
1. Здійснювати обчислення на вузлах, де інформація спочатку була збережена
2. Використовувати обчислювальні потужності систем зберігання
3. Використовувати функціональне програмування для вирішення задач масивно-паралельної обробки
3. Використовувати мову програмування Phyton

8. Методи навчання.

Засвоєння матеріалу забезпечується на лекціях, лабораторних заняттях та самостійній роботі у комп'ютерних класах, обладнаних локальними мережами, Інтернет та програмним забезпеченням. Лекції супроводжуються використанням презентацій та мультимедійного обладнання для полегшення засвоювання матеріалу.

9. Форми контролю.

Контроль знань у слухачів магістерського курсу «Аналітика великих даних» передбачає такі контрольні заходи:

- самоконтроль – є первинною формою контролю знань, який обов'язково забезпечується дистанційним курсом шляхом надання студентам переліку запитань (питань та відповідей на них);
- поточний контроль – здійснюється через систему оцінки безпосередньо викладачем лабораторно-практичних практичних занять та виконаних завдань для самостійної роботи;
- модульний контроль – здійснюється дистанційно в автоматизованому режимі або очному режимі, основною формою якого є тестування;
- підсумковий контроль – це іспит, який складається очно в період призначений деканатом або за індивідуальним графіком, який затверджується навчальним планом. Основною формою підсумкового контролю є тестування.

10.Розподіл балів, які отримують студенти.

Оцінювання студента відбувається згідно положенням «Про екзамени та заліки у НУБіП України» від 27.02.2019 р. протокол № 7 з табл. 1.

Таблиця 1. Співвідношення між національними оцінками і рейтингом здобувача вищої освіти

Оцінка національна	Рейтинг здобувача вищої освіти, бали
Відмінно	90 – 100
Добре	74 – 89
Задовільно	60 – 73
Незадовільно	0 – 59

Для визначення рейтингу студента (слухача) із засвоєння дисципліни $\mathbf{R_{ДИС}}$ (до 100 балів) одержаний рейтинг з атестації (до 30 балів) додається до рейтингу студента (слухача) з навчальної роботи $\mathbf{R_{НР}}$ (до 70 балів): $\mathbf{R_{ДИС} = R_{НР} + R_{АТ}}$.

11.Методичне забезпечення

1. Абдикеев Н.М. Когнитивная бизнес-аналитика Учебник. – М.: ИНФРА-М, 2014. – 511 с.
2. Фрэнкс Билл. Укрощение больших данных М.: Манн, Иванов и Фербер, 2014. – 352 с.

.

12. Рекомендована література

– основна;

1. Гобарева Я.Л., Городецкая О.Ю., Золотарюк А.В. Бизнес-аналитика средствами Excel М.: Вузовский учебник, ИНФРА-М, 2013. – 336 с.
2. Дэвенпорт Том, Хо Ким Джин. О чем говорят цифры. Как понимать и использовать данные Манн, Иванов и Фербер, 2014.
3. Кулешова О.В. Microsoft Excel 2010. Уровень 2. Расширенные возможности М.: Центр компьютерного обучения «Специалист» при МГТУ им. Н.Э. Баумана, 2012. – 91 с.
4. Маккинни У. Python и анализ данных М.: ДМК Пресс, 2015. – 482 с.
5. Осетрова И.С., Осипов Н.А. Microsoft Excel 2010 для аналитиков Учебное пособие. – СПб.: НИУ ИТМО, – 2013. – 65 с.
6. Соммервилл И. Инженерия программного обеспечения, 6-е издание: Пер. с англ. – М.: Издательский дом "Вильямс", 2012. – 624 с.
7. Фрэнк Билл. Революция в аналитике. Как в эпоху Big Data улучшить ваш бизнес с помощью операционной аналитики М.: Альпина Паблишер, 2014. – 430 с.
8. Шаховська Н. Б., Болюбаш Ю. Я. Модель великих даних “сутність-характеристика”. Режим доступу: http://ena.lp.edu.ua:8080/bitstream/ntb/29775/1/20_186-196.pdf
9. National Research Council. 2013. Frontiers in Massive Data Analysis. Washington, D.C.: The National Academies Press
10. Big Data analytics: Future architectures, Skills and roadmaps for the CIO – 2011. – IDC/SAS

– допоміжна.

1. Big Data Visualization: Turning Big Data into Big Insights. The Rise of Visualization-based Data Discovery Tools. White Paper. Intel IT Center. March 2013
2. Big Data: The Next Frontier for Innovation, Competition, and Productivity – 2011. – McKinsey Global Institute
3. Martin Hilbert. Big Data for Development: From Information- to Knowledge Societies", – 2013. – SSRN Scholarly Paper No. ID 2205145). Rochester, NY: Social Science Research Network
4. DJ Patil. Building Data Science Teams. O’Reilly. 2011. ISBN: 978-1-449-31623-5 http://cdn.oreilly.com/radar/2011/09/Building-Data-Science-Teams.pdf

13.Інформаційні ресурси

1. IBM Analytics http://www.ibm.com/analytics/us/en/technology/hadoop/hadoop-trials.html
2. IBM Cloud https://www.ibm.com/cloud-computing/bluemix/?lnk=hp_trials_uauk

3. IBM Bluemix Promo Code - 6 Month Trial https://ibm.onthehub.com/WebStore/OfferingDetails.aspx?o=bb3528b7-2b63-e611-9420-b8ca3a5db7a1
4. Hadoop: Built for big data, insights, and innovation http://www.ibm.com/analytics/us/en/technology/hadoop/
5. IBM BigInsights http://www.ibm.com/analytics/us/en/technology/biginsights/
6. Виктор Маер-Шенбергер, Кеннет Кукьер. Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим. – М.: «Манн, Иванов и Фербер», 2013, 240 с. ISBN 978-5-91657-936-9 http://www.mann-ivanov-ferber.ru/books/paperbook/big_data/
7. Weka Machine learning software to solve data mining problems https://sourceforge.net/projects/weka/?source=typ_redirect
8. Books Ngram Viewer https://books.google.com/ngrams
9. Революция Big Data: Как извлечь необходимую информацию из «Больших Данных»? http://statsoft.ru/products/Enterprise/big-data.php
10. Бесплатные программы для статистического анализа данных http://boris.bikbov.ru/2013/12/01/besplatnyie-programmyi-dlya-statisticheskogo-analiza-dannyih/
11. https://www.r-bloggers.com/
12. Мова програмування R [Електронний ресурс]: https://cran.r-project.org
13. http://r-analytics.blogspot.ru/p/rstudio.html#.WDifOrnzuwI
14. Середовище для розробки програм на R – R Studio [Електронний ресурс]: http://www.r-studio.com
15. http://www.tadviser.ru/index.php/Статья:Большие_данные_(Big_Data)