МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ ІМЕНІ СЕМЕНА КУЗНЕЦЯ

"ЗАТВЕРДЖУЮ"
Заступник керівника
(проректор з науково-педагогічної роботи)
М.В.Афанасьєв

Високопродуктивні системи обробки великих даних

робоча програма навчальної дисципліни

Галузь знань	**12 Інформаційні технології**
Спеціальність	**122 Комп'ютерні науки**
Освітній рівень	**другий (магістерський) рівень**
Освітня програма	**Комп'ютерні науки**

Вид дисципліни	**базова**
Мова викладання, навчання та оцінювання	**українська**

Завідувач кафедри ІС доц.. Ушакова І.О.

Харків
ХНЕУ ім. С. Кузнеця

2019

ЗАТВЕРДЖЕНО
на засіданні кафедри інформаційних систем
Протокол № 1 від 30.08.2019 р.

Розробники:
Мінухін Сергій Володимирович, доктор технічних наук, професор кафедри інформаційних систем

Лист оновлення та перезатвердження робочої програми навчальної дисципліни

Навчальний рік	Дата засідання кафедри – розробника РПНД	Номер протоколу	Підпис завідувача кафедри

1. Вступ

Анотація навчальної дисципліни:

Розвиток технологій розподілених обчислень, а також наявність потужних надпродуктивних систем, які є загальнодоступними для комерційних та науково-дослідних організацій, дозволяють натепер ефективно обробляти дані великих об'ємів. Значний вплив на цей розвиток базується на програмних технологіях побудови масштабованих багатомашинних інформаційно-обчислювальних систем, що забезпечують розподілену та паралельну обробку надвеликих масивів даних. За кордоном сукупність таких технологій позначається терміном Big Data (англ. - великі дані). Умови зростання обсягів даних і збільшення залежності бізнес-процесів підприємств різних галузей від ефективності їх обробки та доступу визначають потреби створення розподілених інформаційних систем різних рівнів, які мають бути представленими у вигляді автономних систем, кластерних рішень тощо Проблема масштабованості при збільшенні інтенсивності даних повинна розв'язуватися сучасними засобами розподілених середовищ у поєднанні з технологіями паралельних інтерфейсів та розподілених файлових систем та сховищ даних. Перелічені завдання розв'язуються на основі розподілених інформаційно-комунікаційних систем.

Основними завданнями дисципліни «Високопродуктивні системи обробки великих даних» є:

формування у студентів компетентностей з принципів функціонування високопродуктивних систем для зберігання та оброблення даних, що організуються за замовленням, для проведення наукових досліджень та задля підвищення продуктивності обчислювального середовища організацій та відповідно ефективності їх бізнес-рішень, що приймаються;

набуття компетентностей щодо вибору певної моделі розподілених обчислень, вибір, встановлення та налаштування програмного забезпечення (фреймвьорків) для роботи у розподілених середовищах на локальному ресурсі та рівні обчислювального кластеру;

набуття компетентностей щодо роботи з великими даними, їх аналізу задля прийняття ефективних управлінських рішень;

поглиблення теоретичних знань, що необхідні для вирішення задач передавання та переробки інформації, автоматизації обробки великих даних у різних предметних областях, а також оволодіння практичними навичками використання та проектування розподілених баз даних та розробки програмних додатків в розподіленому режимі;

встановлення та конфігурування програмного забезпечення та технологій розподіленого зберігання великих даних та отримання практичних навичок роботи з ним для створення та завантаження додатків та розподілених БД в розподіленому режимі.

Мета навчальної дисципліни: Метою викладання навчальної дисципліни "Високопродуктивні системи обробки великих даних" є формування системи теоретичних знань і придбання практичних умінь і навичок з питань використання технологій розподілених та паралельних обчислень на базі технологій високопродуктивних систем.

Курс	**1М**	
Семестр	**1**	
Кількість кредитів ECTS	**5**	
Аудиторні навчальні заняття	**лекції**	**16**
	лабораторні	**24**

Самостійна робота		**110**
Форма підсумкового контролю	**іспит**	

Структурно-логічна схема вивчення навчальної дисципліни:

Попередні дисципліни	Наступні дисципліни
Операційні системи	Хмарні обчислення
Комп'ютерні мережі	
Розподілені та паралельні обчислення	
Програмування	

2. Компетентності та результати навчання за дисципліною:

Компетентності	Результати навчання
Знати сучасні стандарти високопродуктивних систем обробки даних	Визначати особливості та характеристики великих даних для використання високопродуктивних систем
Здатність до вибору сучасних архітектур високопродуктивних розподілених та паралельних обчислювальних систем	Визначати потрібні технології для використання сервісів хмарних платформ для проведення навчальних та наукових досліджень
Здатність застосовувати знання стандартів в області розподілених інформаційно-комунікаційних технологій та систем при виборі певної високопродуктивної системи	Підвищувати ефективність обчислень та обробки великих даних шляхом обґрунтованого вибору архітектури високопродуктивної системи та відповідного програмного забезпечення для її експлуатації
Здатність встановлювати, тестувати та експлуатувати компоненти програмного забезпечення для роботи в високопродуктивних системах	Розробляти ефективні алгоритми щодо створення додатків з використанням розподілених файлових систем засобами високопродуктивних систем
Здатність розв'язувати проблеми масштабованості, проектування та експлуатації розподілених систем для проведення високопродуктивних обчислень	Підвищувати продуктивність обробки великих даних в високопродуктивних системах

3. Програма навчальної дисципліни

Змістовий модуль 1. Стандарти, архітектури та принципи побудови розподілених високопродуктивних систем для обробки великих даних

Тема 1. Основні поняття великих та особливості систем обробки великих даних.

1.1. Поняття та характеристика великих даних.

Історія поняття великих даних. Основні характеристики великих даних. Процеси збору та аналізу великих даних.

1.2. Обробка та аналіз великих даних. Сучасні засоби обробки великих даних

Тема 2. Базові архітектури високопродуктивних систем.

2.1. Особливості побудови процесорів для підвищення продуктивності обчислень. Суперскалярні процесори. Архітектури сучасних ЕОМ для реалізації високопродуктивних обчислень. Векторна обробка даних. Векторні процесори. Організація оперативної пам'яті.

2.2. Зв'язок між елементами паралельних обчислювальних систем. Кластери робочих станцій. Приклад високопродуктивної архітектури на базі системи HP/ConvexExemplar SPP1600. Переваги кластер них паралельних і розподілених обчислень.

Тема 3. Розподілені файлові системи.

3.1. Характеристика розподілених файлових систем.

Призначення та основні принципи побудови файлової системи Google File System.

3.2. Архітектура файлової системи GFS. Склад та функції компонент.

Тема 4. Apache Hadoop.

4.1. Розподілена файлова система Hadoop (HDFS). Архітектура HDFS. HDFS порівняно з NFS.

4.2. Типовий кластер Hadoop - об'єдннаня MapReduce та HDFS. Призначення вузлів NameNode, DataNode.

4.3. Архітектура MapReduce. Компоненти архітектури MapReduce.

4.5. Управління роботами та завданнями.

4.6. Основні функції MapReduce.

Тема 5. Програмна модель MapReduce.

5.1. Операції Map та Reduce.

5.2. Операції . введення і виведення даних та результатів.

5.3. Розробка додатків в програмному середовищі MapReduce.

5.4. Hadoop/MapReduce обмеження.

Тема 6. Принципи організації та функціонування фрейвьорку Apache Spark. Основні компоненти т їх призначення.

6.1. Характеристика та особливості побудови платформи Apache Spark.

Склад та призначення компонент, що підтримуються Apache Spark.

6.2. Характеристика компоненти Core.

6.3. Характеристика компоненти Spark SQL.

6.4. Характеристика компоненти машинного навчання MLLib.

6.5. Характеристика компоненти GraphX.

Тема 7. Принципи роботи архітектури Apache Spark.

7.1. Компоненти та інструменти Spark.

7.2. Архітектура Spark.

7.3. . Інтерфейси для Spark SQL та взаємодії з Spark.

7.4. Життєвий цикл програми Spark.

Тема 8. Режими розгортання Apache Spark.

8.1. Характеристика режиму Local.

8.2. Характеристика режиму Standalone.

8.3. Характеристика режиму Mesos.
8.4. Характеристика режиму Yarn.

Тема 9. Організація роботи Apache Spark.

9.1 Поняття та функції RDD. Приклади програмування на основі RDD.
9.2. Поняття та функції DataFrames.
9.3. Поняття та функції Datasets.

Теми лабораторних робіт.

Лабораторна робота №1.
Установка Spark за допомогою Vagrant.
Лабораторна робота №2.
Установка кластера Apache Spark в автономному режимі.
Лабораторна робота №3.
Установка та налаштування Apache Spark YARN кластера.

4. Порядок оцінювання результатів навчання

Система оцінювання сформованих компетентностей у студентів враховує види занять, які згідно з програмою навчальної дисципліни передбачають лекційні та лабораторні заняття, а також виконання завдань самостійної роботи. Оцінювання сформованих компетентностей у студентів здійснюється за накопичувальною 100-бальною системою. Відповідно до Тимчасового положення "Про порядок оцінювання результатів навчання студентів за накопичувальною бально-рейтинговою системою" ХНЕУ ім. С. Кузнеця, контрольні заходи включають:

поточний контроль, що здійснюється протягом семестру під час проведення лекційних, лабораторних занять і оцінюється сумою набраних балів (максимальна сума – 60 балів; мінімальна сума, що дозволяє студенту складати іспит, – 35 балів);

модульний контроль, що проводиться у формі письмової контрольної роботи як проміжний міні-екзамен з ініціативи викладача з урахуванням поточного контролю за відповідний змістовий модуль і має на меті отримати *інтегровану* оцінку результатів навчання студента після вивчення матеріалу з логічно завершеної частини дисципліни – змістового модуля;

підсумковий/семестровий контроль, що проводиться у формі семестрового екзамену відповідно до графіку навчального процесу.

Оцінювання знань студента під час лабораторних занять та виконання індивідуальних завдань проводиться за наступними критеріями:

здатність підвищувати ефективність обчислень та обробки великих даних шляхом обґрунтованого вибору архітектури певної високопродуктивної системи та відповідного програмного забезпечення для її використання;

здатність застосовувати технології розподілених файлових систем і сховищ даних та розподілених СУБД на базі різних моделей даних для побудови розподілених систем збереження даних;

здатність встановлювати та налаштовувати ПЗ для високопродуктивних обчислень в рамках розподілених систем, у тому числі й на хмарних платформах;

здатність розробляти ефективні алгоритми для створення додатків з використанням розподілених файлових систем та засобів високопродуктивних систем.

Підсумковий контроль знань та компетентностей студентів з навчальної дисципліни здійснюється на підставі проведення семестрового екзамену, завданням якого є перевірка розуміння студентом програмного матеріалу в цілому, логіки та взаємозв'язків між окремими розділами, здатності творчого використання накопичених знань, вміння формулювати своє ставлення до певної проблеми навчальної дисципліни тощо.

Екзаменаційний білет охоплює програму дисципліни і передбачає визначення рівня знань та ступеня опанування студентами компетентностей.

Кожен екзаменаційний білет складається із 3 завдань (ситуаційного, діагностичного та евристичного).

Практичні завдання передбачають вирішення типових професійних завдань фахівця на робочому місці та дозволяють діагностувати рівень підготовки і компетентності студента з навчальної дисципліни.

Результат семестрового екзамену оцінюється в балах (максимальна кількість – 40 балів, мінімальна кількість, що зараховується, – 25 балів) і проставляється у відповідній графі екзаменаційної "Відомості обліку успішності".

Студента слід **вважати атестованим**, якщо сума балів, одержаних за результатами підсумкової/семестрової перевірки успішності, дорівнює або перевищує 60. Мінімально можлива кількість балів за поточний і модульний контроль упродовж семестру – 35 та мінімально можлива кількість балів, набраних на екзамені, – 25.

Підсумкова оцінка з навчальної дисципліни розраховується з урахуванням балів, отриманих під час екзамену, та балів, отриманих під час поточного контролю за накопичувальною системою. Сумарний результат у балах за семестр складає: "60 і більше балів – зараховано", "59 і менше балів – не зараховано" та заноситься у залікову "Відомість обліку успішності" навчальної дисципліни.

Розподіл балів за тижнями згідно з технологічною картою подано в табл. 1.

Таблиця 1

Розподіл балів за тижнями

Теми змістового модуля			Лекційні заняття	Лабораторні заняття	Письмова контрольна робота	Усього
Стандарти, архітектури та принципи організації високопродуктивних систем для	**ТЕМА 1.** Основні поня великих та особливос систем обробки велик даних.	1 тиждень	0,5			0,5
		2 тиждень		1		1
	ТЕМА 3. Розподілені файлові системи. **ТЕМА 2.** Базові архітектури високопродуктивних систем	3 тиждень	0,5			0,5
		4 тиждень		1		1
	ТЕМА 4. Apache Hadoop.**ТЕМА 5.** Програмна модель	5 тиждень	0,5	9		9,5

	MapReduce.					
		6 тиждень		1		1
	ТЕМА 6. Принципи організації та функціонування фрейвьорку Apache Spark. Основні компоненти та їх призначення.**ТЕМА 7.** Принципи роботи архітектури Apache Spark.	7 тиждень	0,5			0,5
		8 тиждень		1	9	10
	ТЕМА 8. Режими розгортання Apache Spark.	9 тиждень	0,5	1		1,5
		10 тиждень		10		10
	ТЕМА 8. Режими розгортання Apache Spark.	11 тиждень	0,5			0,5
		12 тиждень		1		1
	ТЕМА 9. Організація роботи Apache Spark	13 тиждень	0,5	1		
		14 тиждень		1		1
	ТЕМА 9. Організація роботи Apache Spark	15 тиждень	0,5	1	9	10,5
		16 тиждень		10		10
	Іспит					40
	Усього		4	38	18	100

Виставлення підсумкової оцінки здійснюється за шкалою, наведеною в табл. 2.

Таблиця 2

Шкала оцінювання: національна та ЄКТС

Сума балів за всі види навчальної діяльності	Оцінка ЄКТС	Оцінка за національною шкалою	
		для екзамену, курсового проекту (роботи), практики	для заліку
90 – 100	A	відмінно	зараховано
82 – 89	B	добре	
74 – 81	C		
64 – 73	D	задовільно	
60 – 63	E		
35 – 59	FX	незадовільно	не зараховано
1 – 34	F		

5. Рекомендована література

Основна

1. Blackwell M., Sen M. Large Datasets and You. [Електронний ресурс]. – Режим доступу: http://www.mattblackwell.org/files/papers/bigdata.pdf.
2. Ross N. FasteR! HigheR! StrongeR! – A Guide to Speeding Up R Code for Busy People. [Електронний ресурс]. – Режим доступу: http://www.noamross.net/blog/2013/4/25/faster-talk.html.
3. Ryan R. Rosario. Taking R to the Limit, Part II: Working with Large Datasets. [Електронний ресурс]. – Режим доступу: http://www.bytemining.com/wp-content/uploads/2010/08/r_hpc_II.pdf.
4. Big Data Specialization. [Електронний ресурс]. – Режим доступу: https://www.coursera.org/specializations/big-data.
5. Mining Massive Datasets. Onlinecourse. [Електронний ресурс]. – Режим доступу: https://online.stanford.edu/course/mining-massive-datasets-self-paced.
6. Smith M.D., Telang R. Streaming, Sharing, Stealing: Big Data and the Future of Entertainment (MIT Press). 2016. [Електронний ресурс]. – Режим доступу: https://www.amazon.com/Streaming-Sharing-Stealing-Future-Entertainment/dp/0262034794/.
7. Karimi H.A. Big Data: Techniques and Technologies in Geoinformatics. 2014. [Електронний ресурс]. – Режим доступу: https://www.amazon.com/Big-Data-Techniques-Technologies-Geoinformatics-ebook/dp/B00HZNQKMM/.
8. Marz N., Warren J. Big Data: Principles and best practices of scalable realtime data systems. 2015. [Електронний ресурс]. – Режим доступу: https://www.amazon.com/Big-Data- Principles-practices-scalable/dp/1617290343/.

Додаткова

9. Bahga A., Madisetti V. Big Data Science & Analytics: A Hands-On Approach. 2016. [Електронний ресурс]. – Режим доступу: https://www.amazon.com/Big-Data-Science- Analytics-Hands/dp/0996025537/.
10. Marr B. Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance. 2015. [Електронний ресурс]. – Режим доступу: https://www.amazon.com/Big-Data-Analytics-Decisions-Performance/dp/1118965833/.
11. Marr B. Data Strategy: How to Profit from a World of Big Data, Analytics and the Internet of Things. 2017. [Електронний ресурс]. – Режим доступу: https://www.amazon.com/Data.
12. Jones H. Data Analytics: An Essential Beginner's Guide To Data Mining, Data Collection, Big Data Analytics For Business, And Business Intelligence Concepts. 2018. [Електронний ресурс]. – Режим доступу: https://millionbooks.best/downloads/data-analytics-the-ultimate-beginners-guide-to-data-analytics.

Інформаційні ресурси в Інтернеті

13. Високопродуктивні системи обробки великих даних (122 Комп'ютерні науки). Сайт ПНС ХНЕУ ім. С.Кузнеця. - [Електронний ресурс]. – Режим доступу: https://pns.hneu.edu.ua/course/view.php?id=5476.
14. Apache Spark. [Електронний ресурс]. – Режим доступу: https://spark.apache.org.
15. apache-spark-internals [Електронний ресурс]. – Режим доступу: https://www.sigmoid.com/apache-spark-internals/.

16. Spark-ecosystem [Електронний ресурс]. – Режим доступу: http://www.jorditorres.org/spark-ecosystem/.

17. Apache Hadoop YARN: The Nextgeneration Distributed Operating System [Електронний ресурс]. – Режим доступу: https://events.static.linuxfound.org/sites/events/files/slides/ApacheCon%2714%20YARN%20Presentation.pdf.

18. YARN scheduler policies [Електронний ресурс]. – Режим доступу: http://www.corejavaguru.com/bigdata/hadoop-tutorial/yarn-scheduler.

19. YARN Schedulers – FIFO, Fair, and Capacity [Електронний ресурс]. – Режим доступу: http://discuss.itversity.com/t/yarn-schedulers-fifo-fair-and-capacity/18098.

20. Apache_Mesos [Електронний ресурс]. – Режим доступу: https://ru.bmstu.wiki/Apache_Mesos.

21. YARN [Електронний ресурс]. – Режим доступу: https://ru.bmstu.wiki/YARN_(Yet_Another_Resource_Negotiator).