КИЇВСЬКИЙ УНІВЕРСИТЕТ ІМЕНІ БОРИСА ГРІНЧЕНКА
Кафедра комп'ютерних наук і математики
ПРОГРАМА ЕКЗАМЕНУ
з дисципліни
«АНАЛІЗ ТА ОБРОБКА ВЕЛИКИХ ДАНИХ»
курс 5

Спеціальність	122 Комп'ютерні науки
Освітня програма	Інформаційно-аналітичні системи
Форма проведення	**дистанційна письмова робота**

Тривалість проведення **80 хвилин**
Максимальна кількість балів: **40 балів**
Критерії оцінювання: **10 балів – перше (теоретичне) питання, 10 балів – друге (теоретичне) питання, 10 балів – третє (теоретичне), 10 балів четверте (практичне) завдання.**
Процедура проведення дистанційного екзамену:
1.Початок екзамену: в назначений день об 11.00, закінчення в 12.20. В 12.30 викладач приступає до перевірки письмових робіт.
2. Кожен студент отримає індивідуальне завдання. Завдання до екзамену знаходиться у папці "Big Data" - файл "ExamenBD.doc" хмарного сервісу www.dropbox.com, доступ до якого сьогодні мають усі студенти.
3.Для екзамену необхідно виконати 4 питання. По 1-му питанню з кожної з 4 груп питань. Студент вибирає питання з кожної групи за порядковим номером (прізвищем) у класному журналі.
4. Файл з завершеною письмовою роботою у вигляді фото з смартфона (або ж у форматі .doc) студент надсилає до папки "Запити файлів/Екзамен" хмарного сервісу www.dropbox.com для перевірки. Доступ до цієї папки автоматично розсилається усім студентам групи по електронній пошті.

Перелік питань до екзамену відповідно до поданих нижче тем змістовних модулів дисципліни:

1. Визначення та термінологія великих даних.
2. Застосування великих даних і базова модель.
3. Структуровані, неструктуровані та напівструктуровані типи великих даних.
4. Метадані для Великих Даних та їх обробка.
5. Характеристики Великих Даних: об'єм, швидкість, різноманітність, достовірність, корисність.
6. Походження даних
7. Процес здобуття даних
8. Ідентифікація даних
9. Збір і фільтрація даних

10. Витяг даних
11. Перевірка й очищення даних
12. Агрегація та подання даних
13. Візуалізація даних
14. Використання результатів аналізу
15. Поняття аналізу даних для Великих Даних
16. Поняття аналітики даних для Великих Даних
17. Аналітична обробка в реальному масштабі часу OLAP для Великих Даних.
18. Концепція аналітичної обробки OLAP на основі багатовимірного куба.
19. Види архітектур OLAP-систем.
20. Технологія OLTP – обробка транзакцій в реальному масштабі часу.
21. Використання сховища даних для надвеликих масивів даних.
22. Вітрини даних для надвеликих масивів даних.
23. Недоліки традиційних баз даних для зберігання надвеликих масивів даних. NoSQL-бази даних.
24. Життєвий цикл аналітики великих даних
25. Процедура ETL: витяг, перетворення й завантаження великих даних.
26. Поняття кластера в сфері обробки надвеликих масивів даних.
27. Система керування базою даних NoSQL. Реплікація масивів даних.
28. Еталонна архітектура для великих даних (Big Data Reference Architecture)
29. Поняття лямбда-архітектури для аналітики великих даних.
30. Особливості розробки інформаційних систем на базі NoSQL-рішень.
31. Технології реплікації.
32. Особливості роботи розподіленої бази даних, яка застосовує нереляційну модель даних.
33. Шардинг для організації сховищ великих даних.
34. Реплікація для організації обробки великих даних.
35. Режим реплікації "головний-підлеглий
36. Взаємозв’язок шардинга і реплікації
37. Теорема CAP (теорема Брюера) для обробки великих даних
38. ACID - набір властивостей для надійної роботи транзакцій бази даних (атомарність, узгодженість, ізольованість, довговічність)
39. BASE - принцип проектування баз даних на основі теореми CAP
40. Основні поняття обробки великих даних
41. Паралельна обробка великих даних
42. Розподілена обробка великих даних
43. Особливості Hadoop - програмної платформи для організації розподіленої обробки великих даних
44. Поняття обробка робочих завдань в системах аналітики великих даних
45. Пакетна обробка великих даних
46. Транзакційна обробка великих даних
47. Обробка великих даних в пакетному режимі
48. Пакетна обробка за допомогою MapReduce
49. Переваги та недоліки застосування моделі Map Reduce.
50. Технологія Apache Hadoop для організації розподіленого оброблення великих об’ємів даних.

51. HDFS - файлова система для зберігання файлів надвеликих розмірів.
52. Задачі Map і Reduce в обробці великих даних
53. Задача Map
54. Об'єднання даних
55. Розбивка даних
56. Перетасування й сортування даних
57. Задача Reduce
58. Інтерпретація алгоритмів MapReduce
59. Обсяг, погодженість, швидкість (ОПШ) великих даних
60. Огляд технологій зберігання великих даних.
61. Бази даних та моделі даних.
62. Підготовка вихідних даних для аналізу: первинна обробка й візуалізація наявних даних.
63. Процес збору великих даних і проблема якості.
64. Труднощі збору великих даних та шляхи їх вирішення.
65. Адміністративні та етичні проблеми: право власності на дані; контроль даних; керування збором даних; конфіденційність і повторна ідентифікація.
66. Web Mining - видобування знань з Web.
67. Етапи Web Mining.
68. Категорії Web Mining.
69. Методи добування Web-контенту.
70. Добування Web-контенту в процесі інформаційного пошуку.
71. Добування Web-контенту для формування баз даних.
72. Добування Web-структур та подання Web-структур.
73. Оцінка важливості Web-структур.
74. Пошук Web-документів з урахуванням гіперпосилань, кластерізація Web-структур.
75. Дослідницька інформація, етап препроцесінгу, етап добування шаблонів, етап аналізу шаблонів й їхнє застосування.
76. Технологія Opinion Mining та її використання в сучасних інформаційних системах прийняття рішень.
77. Роль машинного навчання в обробленні великих даних.
78. Типи задач машинного навчання: навчання з учителем, навчання без учителя та навчання з підкріпленням.
79. Взаємозв’язок машинного навчання з іншими областями аналітики великих даних.
80. Навчання дерев рішень.
81. Навчання асоціативних правил.
82. Штучні нейронні мережі.
83. Мови Phyton і R у сфері роботи з даними.
84. Вибір рішення для зберігання даних на рівні пакетної обробки.

Приклад індивідуального завдання (екзаменаційного білету):

1. Відмінні характеристики Великих Даних.
2. Методи аналітичної обробки Великих Даних в реальному масштабі часу OLAP. Транзакційна обробка Великих Даних.
3. Типи задач машинного навчання: навчання з учителем, навчання без учителя та навчання з підкріпленням.
4. Задача кластерного аналізу. Використання методу k-means (лінійний та нелінійний) для вирішення задачі кластеризації. Пояснити відмінності цих двох методів.

Екзаменатор Гладун А.Я.

Завідувач кафедри Машкіна І.В.