Министерство науки и высшего образования Российской федерации
Калужский филиал федерального государственного бюджетного образовательного учреждения высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»

РАССМОТРЕНО и **ОДОБРЕНО**

на методическом семинаре кафедры *ИУК4 «Программное обеспечение ЭВМ, информационные технологии»*

Протокол № 51.4/02 от « 23 » ноября 2022 г.
Зав.кафедрой_____/Гагарин Ю.Е./

ВОПРОСЫ К ЭКЗАМЕНУ

по дисциплине «Технологии обработки больших данных»
для студентов групп ИУК4-71Б, 72Б

Знать:

1. Дать определение «Большие данные».
2. Приведите методы и техники анализа Больших данных
3. В каких отраслях применяются Большие данные
4. Напишите преимущества и недостатки использования Больших данных
5. Перечислите принципы работы с Большими данными
6. Какие основные признаки характеризуют Большие данные
7. Объясните, почему Большие данные являются одой из глобальных проблем современности
8. Какие используются алгоритмы, применимые в машинном обучении. В чем основная идея такой формы обобщения, как алгоритм PageRank.
9. Перечислите виды обобщения.
10. Опишите и приведите пример такого вида обобщения, как кластеризация
11. Перечислите и опишите модели на основе признаков
12. В чем заключается принцип Бонферрони. Приведите примеры
13. Что такое технология Hadoop. Какая файловая система лежит в ее основе
14. Опишите концепции и структуру HDFS.
15. Опишите архитектуру Hadoop MapReduce
16. Перечислите преимущества, недостатки и ограничения Hadoop MapReduce
17. Приведите примеры использования технологии Hadoop
18. Для чего предназначена платформа ArenaData
19. Дайте определение Apache Hadoop
20. Опишите принцип работы «Озера данных»

21. Дайте определение «Озеро данных»
22. Перечислите преимущества, недостатки архитектуры «Озера данных»
23. Приведите примеры развертывания «Озер данных» для эффективного использования
24. В чем заключается принцип масштабирования. Приведите примеры, где реализован такой подход.
25. Какие механизмы увеличения скорости аналитической обработки существуют в современных базах данных
26. В чем заключается суть комбинирования моделей. Опишите стратегию параллельной обработки данных
27. Как происходит анализ данных при помощи репрезентативной выборки
28. Какая основная цель процесса сжатия данных. Перечислите основные характеристики процессов сжатия.
29. Опишите обратитимый и необратимый способы сжатия данных. Назовите преимущества и недостатки каждого способа.
30. Какая модель лежит в основе любого способа сжатия данных, и какие алгоритмы применяются в этой модели.
31. Какие существуют методы сжатия данных.
32. Какая основная идея заложена в распределенной файловой системе, и на чем основываются ее подходы.
33. Перечислите преимущества и недостатки схемы распределенной файловой системе
34. Приведите примеры распределенных файловых систем и опишите их
35. Опишите структуру проекта Hadoop
36. Из каких компонентов состоит HDFS
37. В каких случаях следует и не следует использовать Hadoop
38. Перечислите аппаратные требования для развертывания Hadoop
39. Опишите концепции, заложенные при проектировании HDFS
40. Перечислите отличительные черты HDFS по сравнению с другими файловыми системами
41. Опишите три вида узлов Hadoop
42. Перечислите ограничения HDFS
43. Раскройте значение термина «клиент HDFS» и перечислите разрешенные операции клиента
44. Объясните принцип взаимодействия компонентов HDFS
45. Перечислите ограничения архитектуры программной платформы Hadoop
46. Перечислите собственные ограничения платформы Hadoop

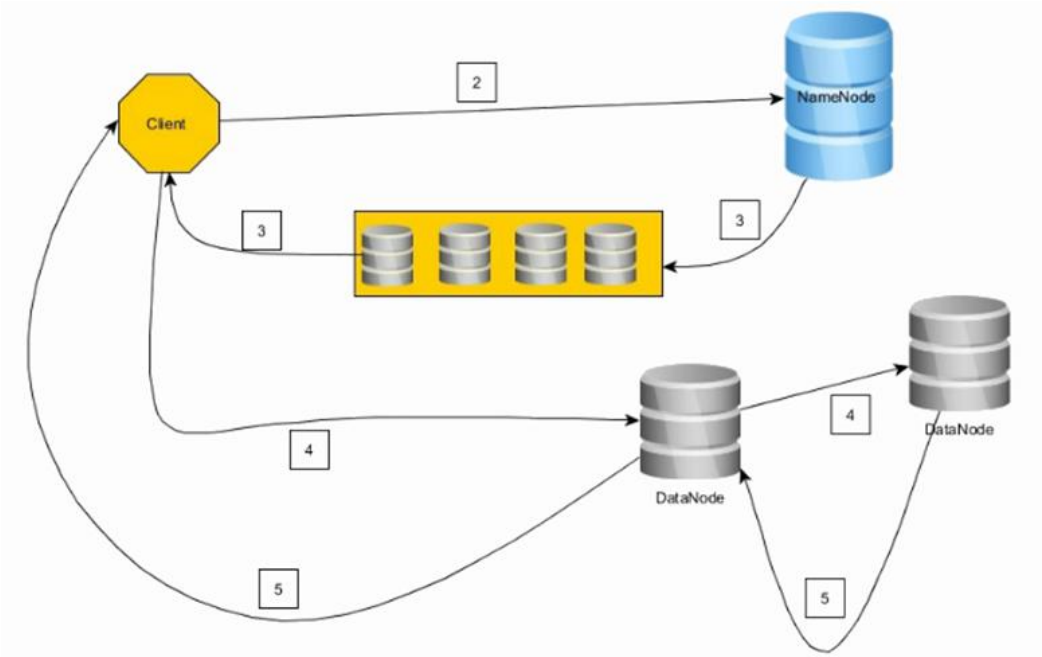
47. Дайте определение термину «порция». В каком виде представлены входные данные распределителей и результаты редукторов. Какой аргумент принимает функция Map, и каким образом он выбираются.
48. Какой аргумент принимает функция Reduce. Опишите принцип работы задачи-редуктора.
49. Дайте определение термину «хэширование», «хэш-функция»
50. Приведите структуру хэш-таблиц
51. Дайте определение коллизии хэш-таблицы. Приведите методы борьбы с коллизиями
52. Дайте определение «коллаборативная фильтрация». Опишите принцип k-расщепления документов.
53. Опишите принцип хэширования шинглов
54. Поясните применение LSH-хэширования для сжатия больших документов
55. Какие проблемы возникают в алгоритме вычисления PageRank
56. Запишите ассоциативное правило «если - то». Каким образом определяются достоверность и поддержка ассоциативного правила
57. Опишите алгоритм с ограниченным числом проходов
58. Раскройте значение термина «кластеризация». Перечислите методы выявления кластеров в данных.
59. Перечислите признаки кластеризации алгоритмов
60. Как устроена иерархическая кластеризация в евклидовом пространстве
61. Какие подходы применяются в принятии решения об остановке процесса кластеризации.
62. Опишите обработку данных в алгоритме BFR
63. Опишите этапы алгоритма CURE
64. Как устроено дерево кластера. Опишите алгоритм GRGPF.
65. Опишите модель потоковых вычислений
66. Дайте определения онлайновым и офлайновым алгоритмам.
67. Дайте определение жадного алгоритма и приведите пример
68. Дайте определение «рекомендательные системы». Какие группы рекомендательных систем можно выделить. На чём основана модель «рекомендательных систем».
69. Опишите структуру матрицы предпочтений. Приведите пример
70. Поняните, что такое TF — IDF, объясните для чего он нужен, и где применяется
71. Опишите и проиллюстрируйте процесс коллаборативной фильтрации.
72. Как можно достичь понижения размерности двух матриц. Раскройте суть UV – декомпозиции.
73. Перечислите основные принципы вычисления среднеквадратичной ошибки

74. Дайте определение термина "социальная сеть". Приведите примеры социальных сетей.
75. Что является метрикой для графов социальных сетей. Какие методы кластеризации применяются в графах. Какие решения проблем иерархической кластеризации можете привести.
76. Дайте определение промежуточности ребра. Приведите пример.
77. Дайте определение полных двудольных графов. Где применяются полные двудольные графы.
78. Опишите методы разрезания на основе собственных векторов. Приведите примеры.
79. Перечислите способы нахождения пересекающихся сообществ. Приведите примеры.
80. Опишите модель графа принадлежности. Раскройте суть метода SimRank.

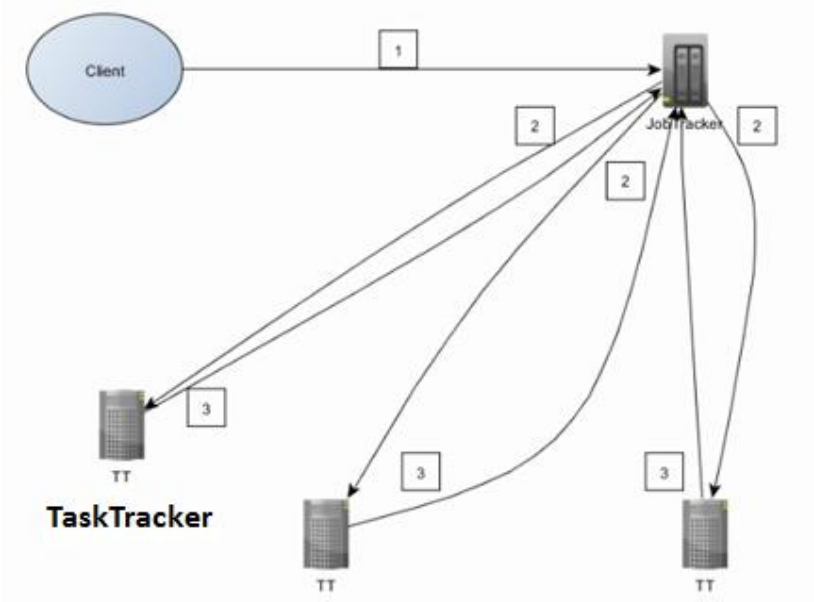
Уметь:

1. Перечислите этапы работы MapReduce
2. Перечислите и опишите основные компоненты Hadoop (yarn, hive, pig, …)
3. Опишите принцип работы Hadoop. В каком виде организованы данные в MapReduce, какие стадии проходит обработка данных
4. Как устроена архитектура «Озера данных». Какие сервисы смогли воплотить эту архитектуру
5. Как устроена архитектура кластерных вычислений. Как решить проблему отказа компонентов такой архитектуры.
6. Опишите схему взаимодействия дистрибутивов HDFS
7. Объясните принцип взаимодействия узлов Hadoop

8. Опишите принцип работы взаимодействия компонентов HDFS по схеме

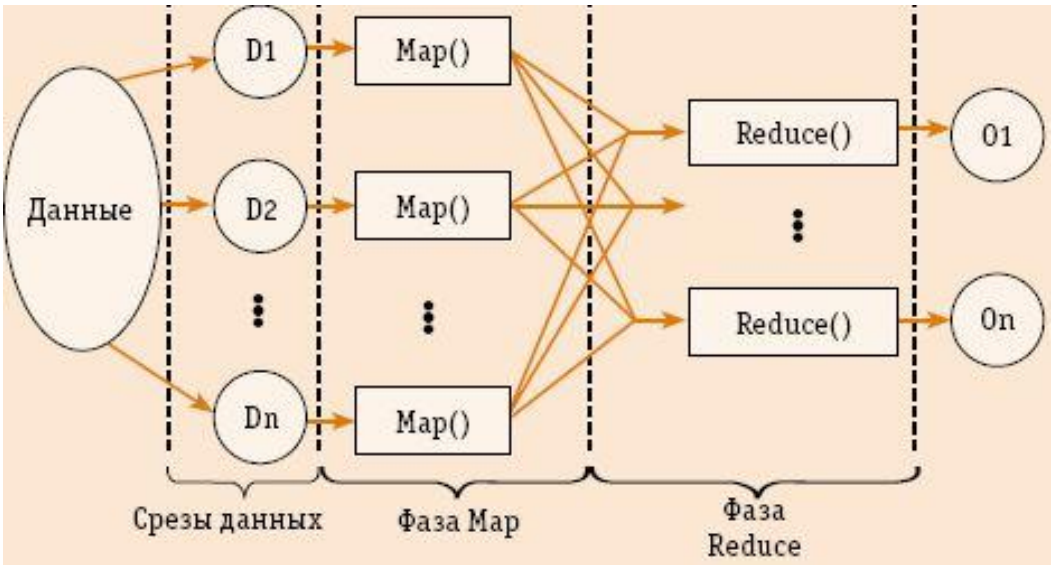


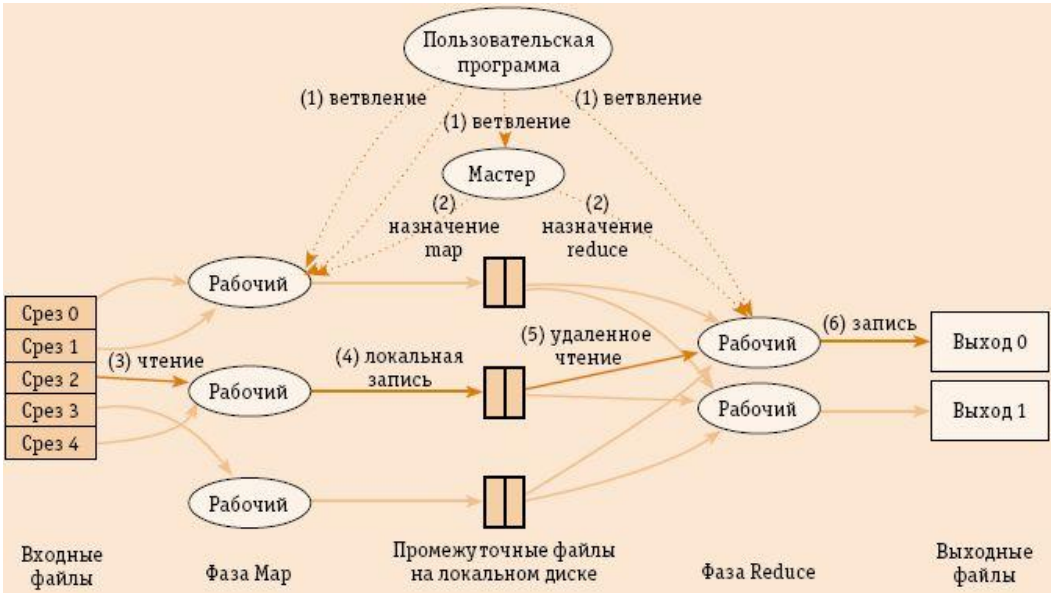
9. Опишите принцип взаимодействия клиента и кластера по следующей схеме:



10. Опишите архитектурное решение для создания файлов

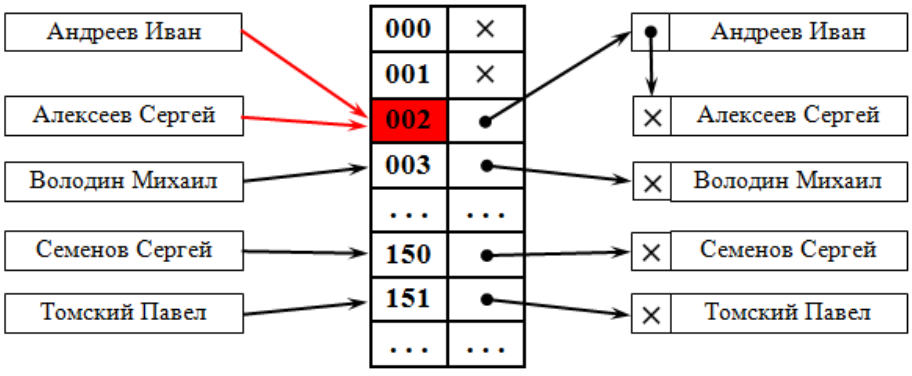
11. Опишите процесс вычислений в MapReduce по схеме:



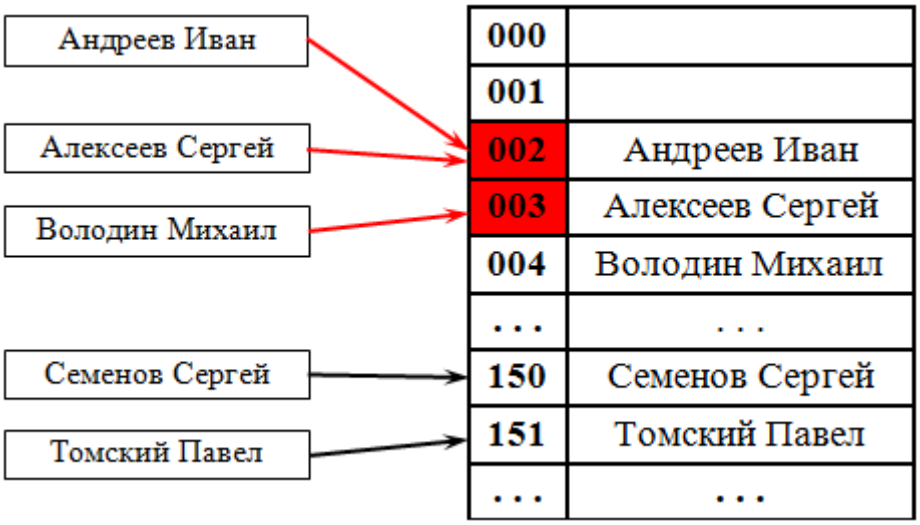


12. Раскройте сущность алгоритма матрично-векторного умножения.

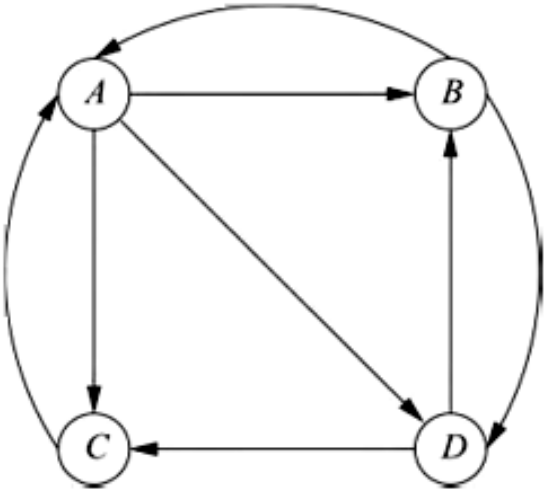
13. Опишите метод цепочек по следующей схеме



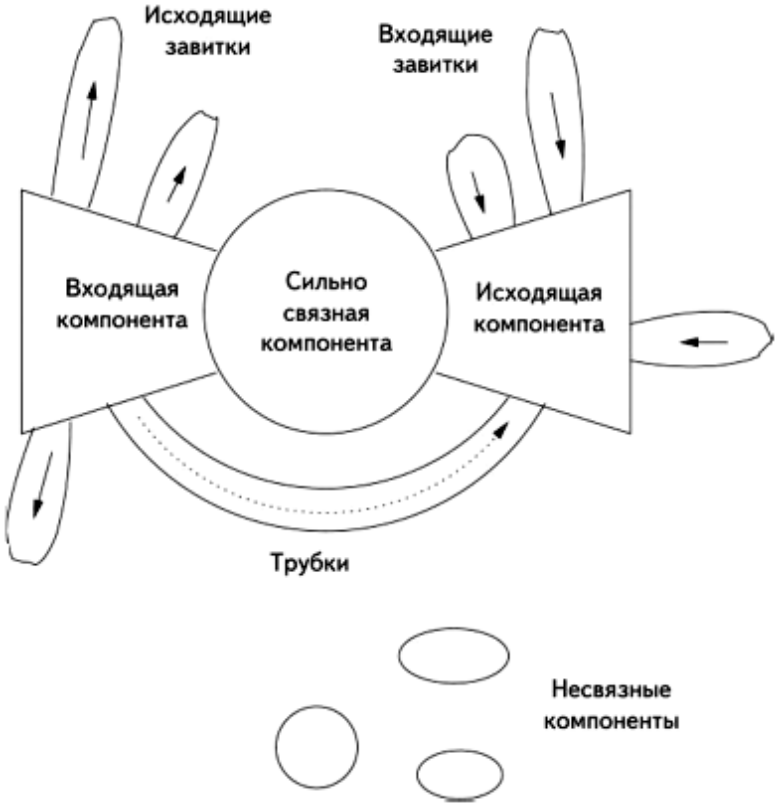
14. Опишите метод открытой адресации по следующей схеме



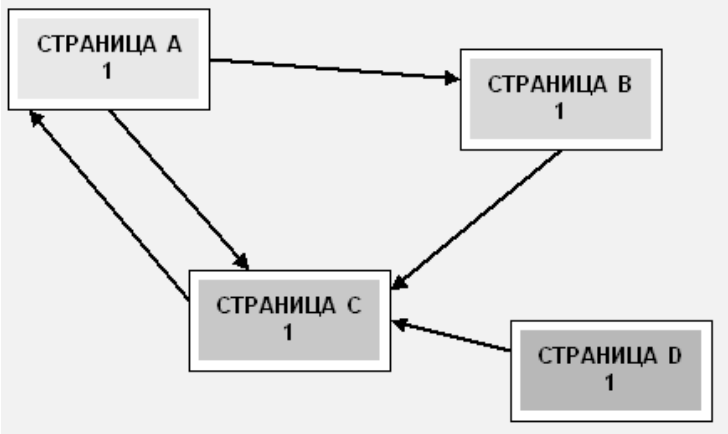
15. Пусть имеется множество точек, называемое **пространством**. **Метрикой** в этом пространстве называется **функция d (x, y)**, которая принимает в качестве аргументов две точки, возвращает вещественное число. Перечислите аксиомы меры расстояний.
16. Докажите, что косинусное расстояние является метрикой
17. Опишите принцип действия функции PageRank, рассматривая веб как ориентированный граф:



18. Опишите принцип двигателя раннего поиска и его компоненты:



19. Опишите принцип работы PageRank, основываясь на знании его веса, используя схему:



20. Опишите модель корзины покупок. Приведите пример этой модели

21. Опишите реализацию Map-Reduce по схеме:

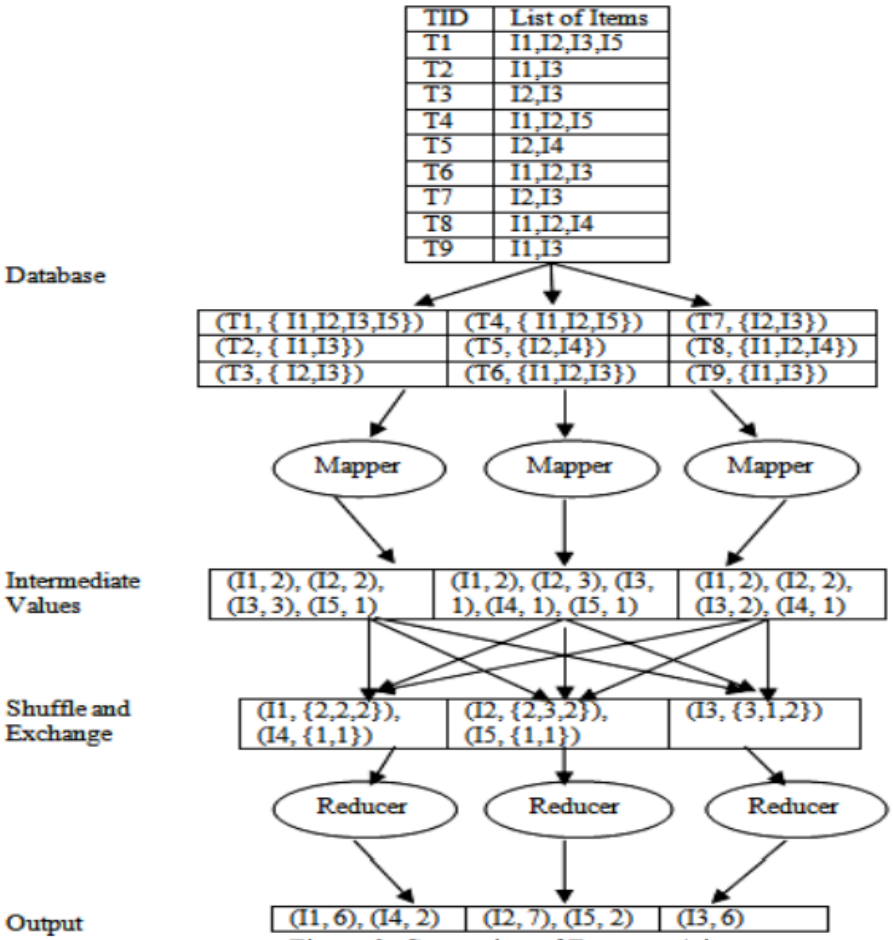
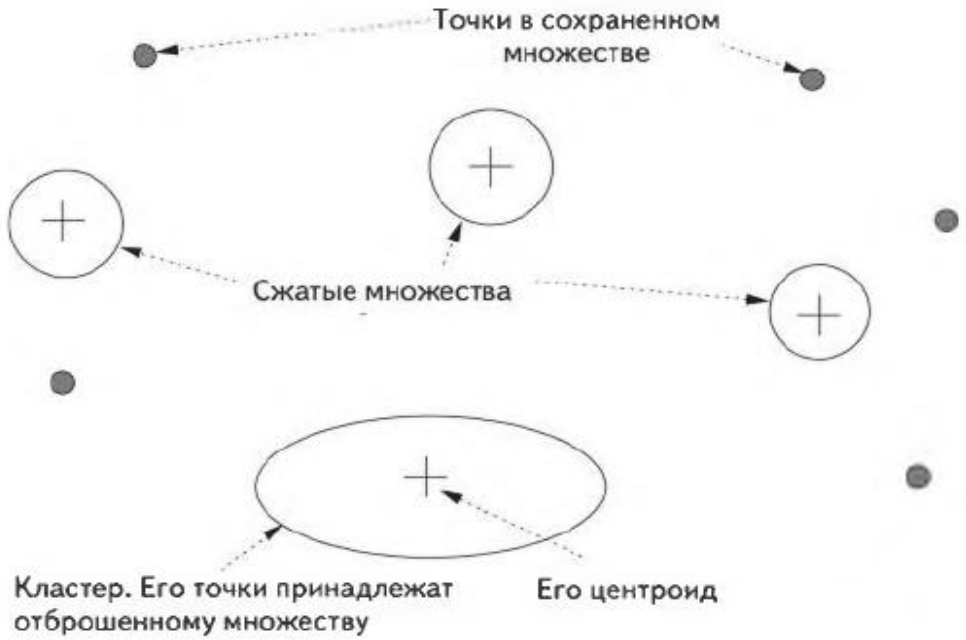


Figure 3: Generation of Frequent 1-itemsets.

22. Классифицируйте алгоритмы кластеризации на основе стратегий
23. Опишите алгоритм Брэдли-Файяда-Рейна

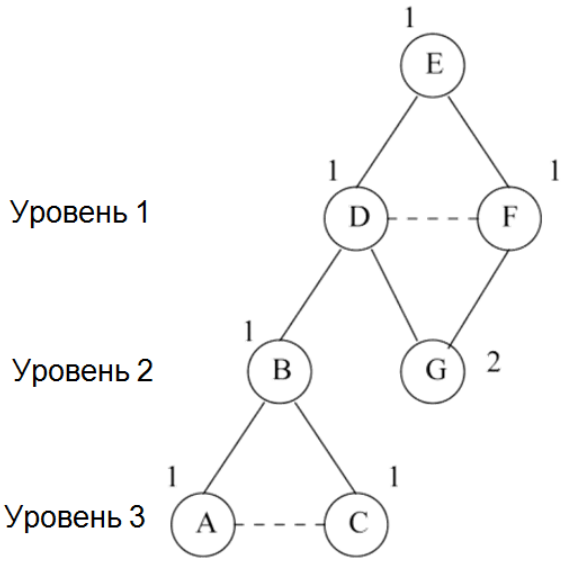


24. Сравните кластеризацию в евклидовых и неевклидовых пространствах
25. Сравните онлайновые и офлайновые алгоритмы
26. Перечислите факты о размерах множеств и паросочетаний.

27. Сравните объясните разницу коэффициентов конкурентоспособности жадного алгоритма паросочетания и алгоритма Balance.
28. Какие модификации необходимо внести, чтобы алгоритм Balance приобрел обобщенный вид:
    Пусть есть два рекламодателя A1 и A2 и один запрос q. Предложенные цены за q и бюджеты показаны в таблице ниже

Рекламодатель	Цена	Бюджет
A1	1	110
A2	10	100

29. Опишите принцип работы алгоритма Гирвана –Ньюмана. По рис 1



30. Напишите алгоритм нахождения сообществ с использованием промежуточности
31. Объясните принцип прямого нахождения сообществ
32. Сравните архитектуры пакетного и оперативного обучения
33. Опишите принцип работы перцептронного классификатора
34. Опишите метод опорных векторов.
35. Перечислите свойства MapReduce
36. Раскройте суть алгоритма Apriori. Перечислите подходы алгоритма Apriori
37. В чём основная идея алгоритма SON. Какие подходы совершает алгоритм Тойвонена
38. Раскройте суть простого рандомизированного алгоритма. Какие подходы применяются. Как предотвратить ошибки в алгоритмах формирования выборки
39. Раскройте суть метода k-средних.
40. Раскройте суть задачи о «максимальном паросочетании»
41. Опишите принцип работы жадного алгоритма для нахождения максимально паросочетания

42. Раскройте суть алгоритма классификации рекомендательной системы.
43. Раскройте суть градиентного спуска
44. Раскройте суть разрезания графов. Напишите основную идею нормализованных разрезов.
45. Раскройте основную идею модели «оценки максимального правдоподобия». Приведите пример

Владеть:

1. Приведите пример использования функции reduce()
2. Решить задачу Word Count и пояснить действия:
   Задача: имеется большой корпус документов. Задача – для каждого слова, хотя бы один раз встречающегося в корпусе, посчитать суммарное количество раз, которое оно встретилось в корпусе.
3. Проиллюстрируйте MapReduce на примере относительно функции Reduce: подсчет количества вхождений каждого слова в коллекцию документов
4. Проиллюстрируйте MapReduce на примере относительно функции Map: подсчет количества вхождений каждого слова в коллекцию документов
5. На рис.1 показан граф с вершинами A B C D. Вставьте значения в таблицу и напишите матрицу переходов.

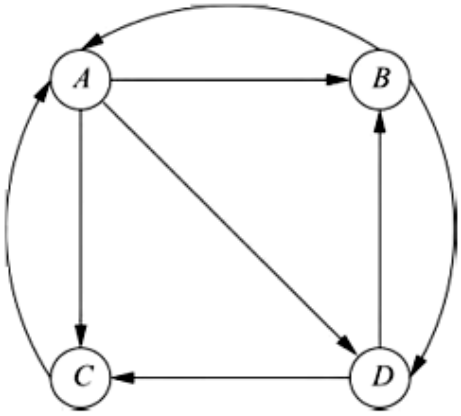
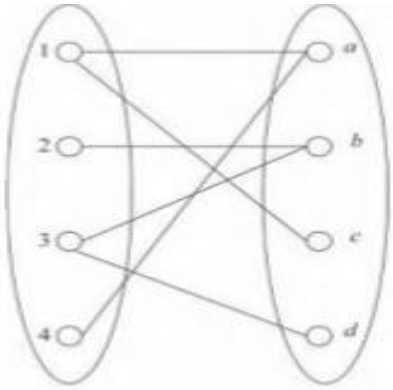


Рис.1

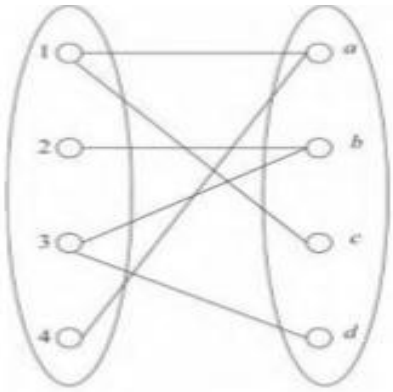
Вершина	Исходящая степень	Последователи
A B C D		

6. Дана формула, определяющая вес PageRank: $PR(A) = (1-d) + d(PR(T_1)/C(T_1) + \ldots + PR(T_n)/C(n1))$. Опишите ее составляющие.
7. Напишите алгоритм инициализации выбора точек в алгоритме k-средних

8. Выпишите множества ребер для совершенного паросочетания двудольного графа рис. 1



9. Для двудольного графа на рис. 1 определить размер паросочетания и обосновать ответ.



10. Решить задачу о ключевых словах :
    Дано:
    1 Множество предложений рекламодателей для поисковых запросов.
    2 Коэффициент кликабельности для каждой пары рекламодатель-запрос.
    3 Бюджеты всех рекламодателей.
    4 Ограничение на количество объявлений, отображаемых в ответ на один поисковый запрос.
    Требуется:
    для каждого поискового запроса вернуть такой набор рекламодателей,
    1 Размер набора не больше ограничения на количество объявлений.
    2 Каждый рекламодатель предложил цену на этот запрос.
    3 У каждого рекламодателя осталось достаточно денег для оплаты щелчка по этому объявлению.
11. Объясните подход жадного алгоритма для онлайновых алгоритмов на примере: Пусть имеется два рекламодателя A и B и всего два возможных запроса x и y. Рекламодатель A торгуется только за x, а B - за x и y. Бюджеты обоих рекламодателей равны 2, и поступила последовательность запросов xxyy.
12. Объясните алгоритм Balance на примере задачи:
    Рекламодатель A торгуется только за x, а B - за x и y. Бюджеты обоих рекламодателей равны 2 Пусть поступила последовательность запросов xxyy.

13. Предположим, что единственными признаками фильма являются актерский состав и средняя оценка. Рассмотрим два фильма с пятью актерами в каждом. Два актера играют в обоих фильмах. Средняя оценка одного фильма равна 3, другого – 4. Как будут выглядеть их векторы. Что показывает масштабный косинус α.
14. Напишите алгоритм бинарного решающего дерева
15. Ниже приведена таблица 1.Объедините все три фильма «фильм1» в один кластер, обозначенный HP, а три фильма «фильм2» –в кластер SW в одну матрицу предпочтений объектов и кластеров объектов. TW – размер объединения

Таблица 1

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

Таблица 2

	HP	TW	SW
A			
B			
C			
D			

16. На рис 1 приведена часть социальной сети. Сущностями являются вершины от A до G. Связь, которую можно условно назвать "друзьями", представлена ребрами. Например, B дружит с A,C и D. Является ли этот граф типичной социальной сетью, те присутствует ли в нем локальность сети? Обоснуйте

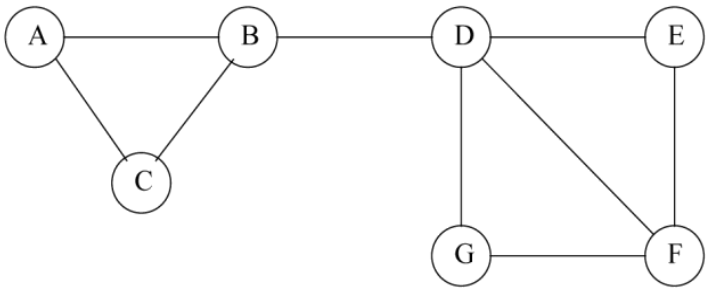


рис 1

17. На рис 1 показан граф. Определить наивысшую промежуточность ребер графа и обосновать.

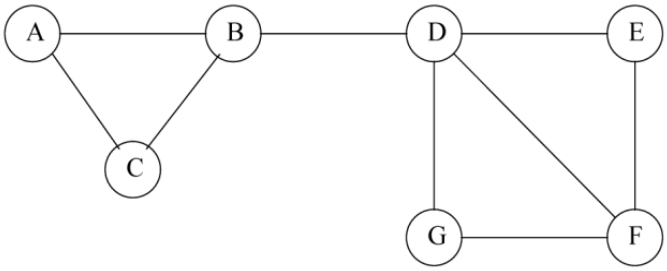


рис1

18. Опишите последовательность нахождения сообществ с использованием промежуточности графа на рис 1

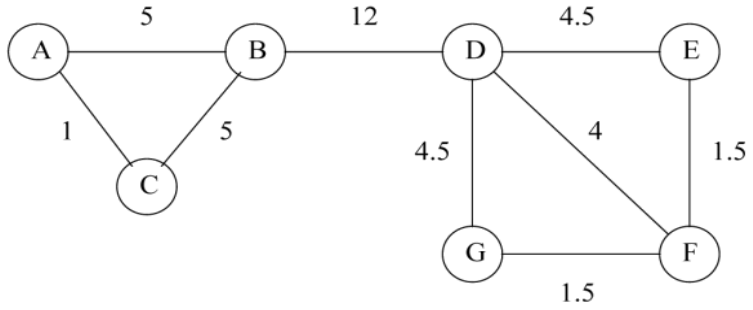


рис 1

19. На рис 1 выберете наилучшее разрезание. Ответ обоснуйте.

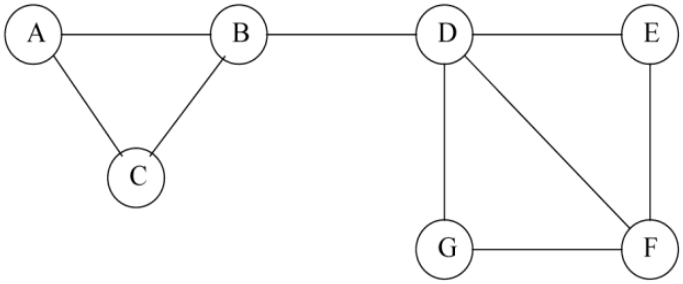


рис 1

20.По графу на рис 1 составьте матрицу смежности.

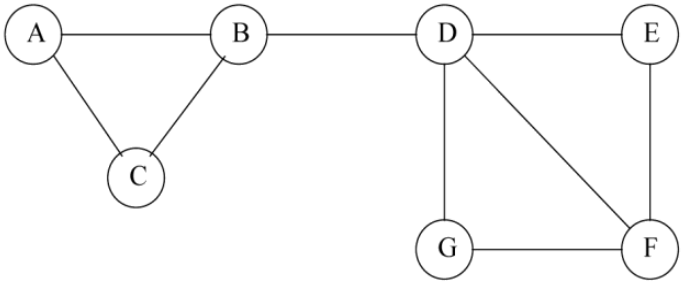


рис1

21. По графу на рис 1 составьте степенную матрицу.

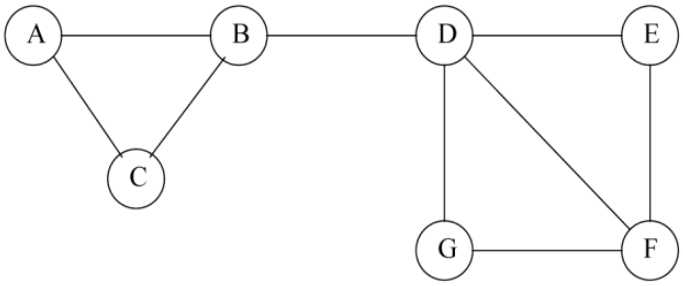


рис1

22. По графу на рис 1 составьте матрицу Лапласа

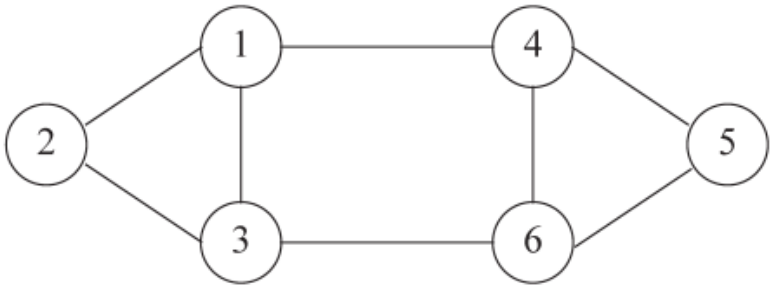


рис 1

23. На рис 1 приведена простая сеть, содержащая три картинки и две ассоциированных с ними метки – «Небо» и «Дерево». Написать матрицу переходов.

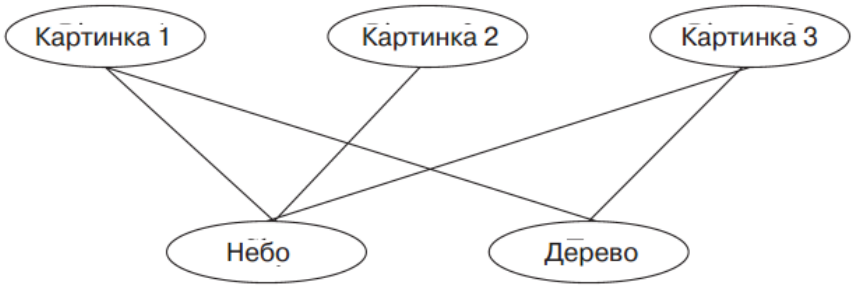


рис 1

24. На рис 1 приведена простая сеть,M – матрица этой сети рис 2 и β = 0.8. Предположим также, что вершина N – Картинка 1, т.е. мы хотим вычислить сходство Картинки 1 с другими картинками. Напишите рекуррентное уравнение для нахождения нового значения v ′:

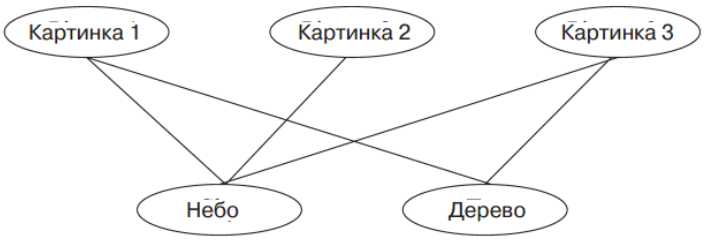


Рис 1

$$\begin{bmatrix} 0 & 0 & 0 & 1/3 & 1/2 \\ 0 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/3 & 1/2 \\ 1/2 & 1 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \end{bmatrix}$$

рис 2

25. Дана матрица на рис 1. Напишите систему уравнений, единичный собственный вектор. Постройте матрицу собственных векторов.

$$M = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix}.$$

рис 1

26. Напишите команду форматирования файловой системы HDFS.
27. Напишите команды запуска служб, необходимых для запуска Hadoop.
28. Напишите команду создания входных файлов в локальной файловой системе
29. Напишите алгоритм решения задачи, используя MapReduce:
    Подсчитать количество строк в файле. Результат должен быть сохранен в файле в виде:
    lile_name lines_count