**БЛОК 1. Фундаментальные знания**

|  |  |
| --- | --- |
| Дана хеш-функция ч mod 6 с открытой адресацией(open addressing) и пустая хеш-таблица с индексами от 0 до 5. Какую позицию в хеш-таблице будет иметь число 29 после вставки туда последовательных чисел 7,11,15,17,27,33? | A.4  ***B.3***  C.2  D.0  E.1 |
| Какая из задач сложнее О(1) для отсортированного массива? | A. поиск значения i-го элемента массива (i известно, i<=N)  ***B.вставка элемента***  C.поиск медианы  D.поиск максимального значения  E. |
| Сколько различных бинарных деревьев можно собрать из трех одинаковых вершин? | A.4  B.2  C.3  D.5  ***E.1*** |
| Какую структуру данных лучше использовать для решения задач проверки баланса скобок в алгебраическом выражении? | A.массив  B.очередь  C.хеш-таблица  D.В-дерево  ***E.стек*** |
| Стек – абстрактный тип данных, представляющий из себя список элементов, организованных по принципу | A. FIFO (first in – first out)  B. LILO (Last in – Last out)  C.FILO (first in - Last out)  ***D. LIFO (Last in- first out)*** |
| Вам нужно написать алгоритм, который подсчитывает kn, где n – неотрицательное целое число. Какова сложность алгоритма для наивного подхода | ***A. О(n)***  B. O (Log)  C.O(1)  D. O(n2) |
| Выберите верные утверждения о верной сортировке | ***A. в худшем случае скорость равна O(n\*2)***  ***B.хорошо подходит для массивов имеющих случайный порядок***  C. в лучшем случае скорость равна O(n^2)  D.хорошо подходит для сортировки почти сортированных массивов |
| Выберите верные утверждения о сортировке кучей | ***A.хорошо подходит для случаев, когда объем памяти ограничен***  B.хорошо подходит для сортировки почти сортированных массивов  ***C***. ***в худшем случае скорость равна O(n\*2)***  D. хорошо подходит для случая, когда нет ограничений по памяти |
| Выберите верные утверждения о сортировке слиянием | A. хорошо подходит для сортировки почти сортированных массивов  B.неплохо параллелится  ***C. в среднем скорость сортировки равна О(n Log n)***  D. . хорошо подходит для случая, когда нет ограничений по памяти |
| Выберите верные утверждения о сортировке вставками | ***A. хорошо подходит для массивов, имеющих случайный порядок***  B. хорошо подходит для сортировки почти сортированных массивов  C. хорошо подходит для больших массивов  ***D. хорошо подходит для небольших массивов*** |
| Какой шаблон проектирования предоставляет интерфейс для создания взаимосвязанных или взаимозависимых объектов без указания их конкретных классов? | A.одиночка (singleton)  B.посредник (mediator)  C.прототип  ***D.абстрактная фабрика*** |
| Для чего используются приватные методы классов? | A.для обработки персональных данных пользователей  ***B.для обеспечения принципа инкапсуляции***  C.является частью интерфейса класса  D.для изменения значений атрибутов объекта пользователем  ***Е. для обеспечения корректной работы внутренней логики объекта*** |
| Объект является экземпляром …. | A. метода  B.параметра  C.приложения  ***D.класса*** |
| По какому принципу ООП соответствует стремление предоставить объект минимальным набором полей и методов необходимых для решения задачи? | A. полиморфизм  B.инкапсуляция  C.наследование  ***D.абстракция*** |
| Как называется определение одной функции в теле другой с целью наследования ее внутреннего контекста? | A. композиция  B.агрегация  ***C.вложение***  D.замыкание |
| В задаче линейного программирования система ограничений должна определять область, представляющую собой: | A. сферу  B.выпуклый или вогнутый многоугольник  C.куб  ***D.выпуклый многоугольник*** |
| При проверке статистической гипотезы, ошибкой первого рода называют | A. отклонение альтернативной гипотезы, которая в действительности является верной  B.принятие нулевой гипотезы, которая в действительности является неверной  C.принятие альтернативной гипотезы, которая в действительности является неверной  ***D.отклонение нулевой гипотезы, которая в действительности является верной*** |
|  |  |

**Блок 2. Ключевые компетенции**

|  |  |
| --- | --- |
| Почему производительность современных аналитических баз данных сильно зависит от объема оперативной памяти? | ***A. данные кашируются в оперативной памяти, что ускоряет их выборку по сравнению с чтением с диска***  B.данные всегда хранятся в оперативной памяти  C. база данных может выполнить JOIN только тех таблиц, которые целиком помещаются в оперативной памяти  D.колоночные базы потребляют больше памяти, чем строковые |
| Какие типы JOIN существуют? Выберите все подходящие варианты | A.SPLIT JOIN  ***B.HASH JOIN***  ***C.MERGE JOIN***  D.DIRECT JOIN |
| Какой из приведенных типов JOIN потребует меньше памяти? | A. зависит от количества полей в таблице  B. HASH JOIN  ***C.MERGE JOIN***  D. зависит от полей , по которым происходит JOIN |
| В каком случае запрос может получить данные из таблицы без обращения к самой таблице? | A. только если на таблице есть индекс по всем полям  B.ни в каком  C.только если на таблице есть primary key  ***D.если на таблице есть индекс по искомому полю*** |
| Какие условия требуются , чтобы был возможен MERGE JOIN? Выберите все подходящие варианты | A. один из соединяемых наборов должен помещаться в память  ***B.наборы данных должны быть отсортированы по всем полям, используемым для JOIN***  C.тип данных в ключе должен быть строковым  D.наборы данных должны быть отсортированы по одному из полей , используемых JOIN |
| Поддерживают ли hash-индексы поиск по нескольким полям, входящих в индекс? | A. да, если поиск идет по последним полям  B. да, если поиск идет по первым полям  ***C. только если поиск идет по всем полям, входящим в индекс***  D. да, если поиск идет любому подмножеству полей |
| Поддерживают ли btree-индексы поиск по префиксу? | A. да, если поиск идет по последним полям  B.только если поиск идет по всем полям, входящим в индекс  C. да, если поиск идет любому подмножеству полей  ***D. да, если поиск идет по первым полям*** |
| Производится выборка данных из двух таблиц(t1и t2) с операцией JOIN по ключу user\_id. Какое из выражений соответствует выбору всех значений user\_id (с повторами или без), которые есть в таблице t2, но нет в таблице t1? Выберите все подходящие варианты | A.t1 join t2 on t1.user\_id=t2.user\_id where t2.user\_id is not null  B. t1 left join t2 on t1.user\_id=t2.user\_id where t1.user\_id is not null  C. t1 left join t2 on t1.user\_id=t2.user\_id where t2.user\_id is not null  ***D. t1 right join t2 on t1.user\_id=t2.user\_id where t1.user\_id is not null*** |
| Предположим, что в таблице в колонке total\_value лежат значения(1,2,NULL,4) Каким будет результат выражения sum (total\_value)при выборке данных из всей таблицы без дополнительных условий? | A. 0  B.Undefined  C.NULL  ***D.7*** |
| Предположим, что в таблице t есть две колонки: department и city. Каким выражением можно получить все уникальные комбинации значений department и city, встречающиеся в строках t1? Выберите все подходящие варианты | ***A. select department, city from t group by department city***  B. select distinct (department), distinct (city) from t  ***C. select distinct(department city) from t***  D. select distinct (department), distinct (city) from t group by department, city |
| Предположим, что в таблице t есть две колонки: department и revenue. Каким выражением можно получить все department по которым сумма revenue больше 30. Выберите все подходящие варианты | A. select department from t where sum (revenue)> 30 group by department  B. select department from t where sum (revenue)> 30 group by revenue  C. select department from t group by department where sum (revenue)> 30  ***D. select department from t group by department having sum (revenue)> 30*** |
| Каким условием можно из таблицы выбрать все записи, начинающиеся на “abc”? Выберите все подходящие варианты | A. like’% abc’  ***B. like’ abc%’***  C. like’(.+) abc’  D. like’ abc(.+)’ |
| Что вернет выражение “drop table t”, если таблицы t не существует? | A.0  B. ничего  ***C.ошибку***  D.NULL |
| Предположим, что в таблице t есть две колонки: department и revenue. Каким выражением можно получить максимальное значение revenue по каждому department (возможно с повторением)?. Выберите все подходящие варианты | A. select revenue from t order by department  ***B. select max (revenue) from t group by department***  ***C. select max (revenue) over (partition by department) from t***  D. select max (revenue) from t order by department) |
| Чем схема базы данных “звезда” отличается от “снежинки” | A. степенью нормализации  B.наличием внешних ключей  ***C.наличием таблиц фактов***  D.количеством хранимых данных |
| Укажите результат выражения в Scala: Option(3).map(\_=>None) | A. Some(3)  B. Option(None)  C. Some(None)  ***D. None*** |
| Какую роль в хранилище данных играют витрины данных (data marts)? | A. хранение нормализованных фактов  B.предоставление связей между фактами и атрибутами  ***C.предоставление данных в форме удобной для внешних систем пользователей***  D. хранение серых данных |
| Для использования собственных классов в MapReduce-заданиях на Java необходимо реализовать минимальный интерфейс(ы): | ***A. Serializable + Comparable***  B. Serializable  C. Writable  D. WritableComparable |
| В чем преимущество подхода ELT по сравнение с ETL? | A.экономия занимаемого дискового пространства  B.экономия процессорных ресурсов  ***C.возможность повторной обработки сырых данных***  D. скорость выполнения задач по загрузке |
| Как в Python можно удалить файл? Выберите все верные варианты | A. os.delete(filename)  ***B. os.remove(filename)***  C. os.rm(filename)  ***D.os.unlink(filename)*** |
| В чем преимущества использования OLAP кубов для BI? | A. возможность получать ответ напрямую из сырых данных  ***B.возможность быстро агрегировать метрики по интересующим изменениям***  C.историчность данных  D.оптимизация занимаемого объема хранилища |
| Предположим, необходимо описать структуру хранилища данных для продуктового магазина. Укажите сущности, которые должны соответствовать таблицам измерений. Выберите все подходящие варианты | A. покупка товара  B.страна производства товара  ***C.количество купленных товаров***  ***D.категория товара*** |
| Как в Python можно удалить файл? Выберите все верные варианты | A. os.delete(filename)  ***B. os.remove(filename)***  C. os.rm(filename)  ***D.os.unlink(filename)*** |
| Для использования собственных классов в MapReduce-заданиях на Java необходимо реализовать минимальный интерфейс(ы): | ***A. Serializable + Comparable***  B. Serializable  C. Writable  D. WritableComparable |
| Укажите результат выражения в Scala: Option(3).map(\_=>None) | A. Some(3)  ***B. Option(None)***  C. Some(None)  D. None |
| Есть файл размером 1.2ГБ. размер блока на HDFS 512 МБ. Replication factor = 2. После записи этого файла на HDFS будет | A. 3 блока. На дисках будет занято 1,2ГБ х 2  ***B. 6 блоков. На дисках будет занято512 МБ х 2***  C. 3 блока. На дисках будет занято 1,2ГБ х 32  D. 2 блока. На дисках будет занято 512 МБ х 2 |
| При записи данных на HDFS | A. клиент пишет данные на NameNode. DataNodes забирают данные с NameNode  B. клиент пишет данные на NameNode. NameNode пишет данные на DataNodes  ***C. клиент пишет данные на DataNodes и сообщение на NameNode, куда записаны данные***  D. NameNode говорит куда писать данные, клиент пишет данные самостоятельно на DataNodes |
| Replication factor HDFS можно менять | A. только для папки и файла  B.только для папки и файла если новый Replication factor больше исходного  ***C.для HDFS, папки, файла***  D.только для папки и файла если новый Replication factor меньше исходного |
| Есть папка с gzip –архивами(compression level =9) лог-файлов. Количество файлов -100. Размер файлов – 100 МБ, размер блока на HDFS – 256 МБ. При запуске MR – задачи сколько mapper–ов будет запрещено? | A. 400  B.100  ***C.нельзя точно определить***  D. 50 |
| Размер блока для существующего файла можно менять: | A.можно менять как угодно  B.можно только если новый размер блока меньше исходного  C.нельзя  ***D.можно только если новый размер блока больше исходного*** |
| В каких случаях в Hive-запросах стоит использовать SORT BY вместо ORDER BY? | A. при работе с большими датасетами, поскольку в SORT BY используется только один редьюсер  B. при работе с большими датасетами, поскольку в SORT BY используется только один маппер  C. при работе с большими датасетами, поскольку в SORT BY используется несколько мапперов  ***D. при работе с большими датасетами, поскольку в SORT BY используется несколько редьюсеров*** |
| Дан кластер из 10 нод, на каждой доступно 16 CPU и 64 GB RAM. Какая из приведенных конфигураций Spark-приложения будет наиболее опимальной? | A. num-executors 160 –executor-cores 1 –executor-memory 4 GB  B. num-executors 40 –executor-cores 4 –executor-memory 16 GB  C. num-executors 39 –executor-cores 4 –executor-memory 12 GB  ***D. num-executors 10 –executor-cores 16 –executor-memory 64 GB*** |
| Какие из утверждений про оптимизатор Catalyst верны? | A.основан на принципах функционального программирования  ***B.имеет Python API для разработчиков***  ***C.использует статистические данные для выбора последовательности применения фильтров к данным***  D. производит вычисления на нодах кластера  Е. формирует физический план выполнения вычислений |
| Что из перечисленного нужно сделать с dataframe, чтобы к нему можно было обращаться при помощи SQL-выражений вида spark.sql(“SELECT\*FROM people”)? | A. ничего не нужно делать SQL-выражения сразу доступны для работы с dataframe  B. dataframe нужно конвертировать в dataset  ***C. из dataframe нужно создать TepmView***  D. dataframe нужно конвертировать в RDD |
| Какие из перечисленных методов Spark Dataframes являются трансформациями (transformations)? | A.withColumnRenamed  B.take  ***C.randomSplit***  ***D.select***  Е. Show  F. count  G. withColumn  H.drop  I. collect  ***J. pivot*** |
| Какие утверждения об Apache Spark верны? | A. каждый контейнер Spark самостоятельно формирует собственный физический план расчетов с учетом того , где относительно него располагаются данные  ***B.концепция PDD позволяет восстановить определенное состояние данных в случае сбоя одного из узлов кластера***  C. использование Broadcast Hash Join позволяет избегать лишнего shuffle за счет копирования одного из датафреймов на все вычислительные узлы  D.Apache Spark – распределенная база данных для realtime-аналитики и машинного обучения  ***E.Shuffle – важный этап графа вычислений Apache Spark, значительно ускоряющий процесс обработки данных*** |
| Укажите результат выполнения следующего набора команд в ОС Linux: false|| true && echo $? | A. 0  ***B.пустой вывод***  C.1  D.ошибка |
| Укажите результат выполнения следующего набора команд в ОС Linux: ps:| gzip>ps.list | ***A. список процессов будет перемещен в сжатый файл с именем ps.list***  B. на экран будут выведены список процессов и информация из сжатого файла с именем ps.list  C. список процессов будет выведен на экран, информация о процессе gzip будет помещена в текстовый файл с именем ps.list  D.информация о процессе gzip будет помещена в текстовый файл с именем ps.list |
| Укажите результат выполнения следующего набора команд в ОС Linux: echo “The firm has fixed costs of $100” | A. The firm has fixed costs of 100  B. “The firm has fixed costs of $100”  ***C. The firm has fixed costs of $100***  D. The firm has fixed costs of 00 |
| Администратор выполняет копирование файлов с использованием команды scp. Необходимо сохранить вывод ошибок копирования в отдельный файл error.log. Выберите правильное действие | A. scp-vr [user@server:/mnt/backup./<2](mailto:user@server:/mnt/backup./%3c2) error.log  B. scp-vr [user@server:/mnt/backup./1](mailto:user@server:/mnt/backup./1)> error.log  C. scp-vr [user@server:/mnt/backup./2](mailto:user@server:/mnt/backup./2)> error.log  ***D. scp-vr*** [***user@server:/mnt/backup./>>***](mailto:user@server:/mnt/backup./%3e%3e) ***error.log*** |
| Укажите результат выполнения следующего набора команд в ОС Linux: DATE=’date’;echo “Today is $DATE” | A. “Today is date”  B. “Today is ‘date’”  C. Today is Fri Dec 13 12:00:00 MSK 2019  ***D. Today is date*** |

**БЛОК 3. Опциональные компетенции**

|  |  |
| --- | --- |
| Выберите верные утверждения про XML: | ***A.есть возможность валидации***  ***B.можно использовать только если обрабатываемый файл – валидный XMK- документ***  ***C.хорошо сжимается***  D.при сохранении на HDFS используется колоночный доступ к содержимому проиндексированных полей |
| Верные утверждения про CSV файлы | ***A.постоянная схема данных***  B.в начале должен быть заголовок (header)  ***C. хорошо сжимется***  D.человекочитаемый |
| Json - это | ***A. содержит схему, и данные***  B.можно использовать только если обрабатываемый файл – валидный json объект (массив)  C.человекочитаемый формат  ***D.постоянная схема данных*** |
| Какой вид авторизации из перечисленных по стандарту корректнее всего использовать в RESTful – сервисах? | A.Cookie-based  ***B.токен в заголовках запроса***  C.HTTP Basic  D. токен в теле запроса |
| Какой HTTP-код возвращается клиенту при успешном удалении ресурса DELETE-запросом? | A. 204 No Content  ***B.200 OK***  C.202 Accepted  D.201 Created |
| Укажите неверное утверждение о REST API: | A. чистой практикой REST является указание типа CRUD-операции в теле URL  B. в случае REST допустимо создавать асинхронные отложенные задачи  C. формат сериализации возвращаемого ресурса может определяться заголовком в теле запроса  ***D. для фильтрации коллекции допустимо использовать GET –параментры в запросе*** |
| Какое поведение ожидается от распределенной базы данных типа СР (в терминах САР теоремы Consistency, Availability, Patrtition tolerance) в случае проблем с сетью? Примечание : Под «корректным » подразумевается ответ, который содержит одни и те же данные вне зависимости о того, к какой ноде отправлен запрос. | A.любой успешный запрос возвращает корректный ответ  ***B.любой запрос возвращает корректный ответ***  C. любой запрос возвращает некорректный ответ  D. любой запрос возвращает ответ |
| Какие существуют архитектуры систем обработки данных? Выберите все подходящие ответы | A. Delta  B.Alfa  ***C.Kappa***  ***D.Lambda*** |
| Чем отличаются аналитические MPP-базы (Vertika, Greenplum) от обычных реляционных баз (PostgreSQL) | A. наличием оптимизатора запросов  B.отсутствием транзакций  C.наличием репликации  ***D.паралельным выполнением одного запроса на нескольких узлах*** |
| Предположим вы случайно некоторое время назад залили обычный Git-репозиторий большой бинарный файл. Осознав проблему , вы теперь пытаетесь удалить этот файл из дерева и из кэша в принципе. Какая команда поможет с этим? | ***A. git revert***  B. git filter-branch –tree- filter  C. git resert  D. git annex drom my\_big\_file |
| Как откатить только что сделанный коммит, не потеряв изменений? | ***A. git resert--soft***  B. git push HEAD^1  C. git revert  D. git resert--hard |
| Допустим вам необходимо выкинуть из дерева несколько коммитов, при этом сама ветка уже ушла вперед. Как это можно сделать, не добавляя новых коммитов? | ***A. git checkout -b***  B. git revert  C. git rebase -i  D. git stash |
| Отметьте все верные типы контейнеров в MS SSIS | A.foreach loop container  ***B.task host container***  C. for loop container  ***D. sequence container*** |
| Что является предопределенным событием (predefined event) в informatica? | A. это событие , которое ожидает появления любого файла в определенном месте  B. это событиен, которое происходит по расписанию  C.это, событие , которое происходит в случайный момент времени  ***D.это событие , которое ожидает появления определенного файла в определенном месте*** |
| Какой из перечисленных ниже типов ключей SAS DI содержит 1 и более колонок, которые соотносятся с уникальным или первичным ключом другой таблицы | A.generated key  B. surrogate key  C.retained key  ***D.foreign key*** |
| Что из перечисленного НЕ является рутиной в IBM DataStage? | A. action function  B.job control routine  C.before/after subrotine  ***D.transform function*** |
| Вы решили, что можно разделить часть ваших пайплайнов на составляющие . Общую логику можно вынести в отдельный DAG. Который будет запускаться из других пайплайнов с помощью оператора –триггера. Какой schedule\_interval надо установить такому DAG (с общй логикой) при его определении? | A. не важно какой schedule\_interval, главное добавить параметрт external\_trigger=True в конструктор  B.поставить startdate в будущее с помощью datatime+timedelta  C.убрать вообще schedule\_interval, но поставить start\_date далеко в будущее  ***D.определить его с schedule\_interval=None***  E. определить DAG с любым schedule и просто явно поставить на паузу |
| Определенный пайплайн сильно грузит источник, из которого читает данные и иногда не успевает отобрать до следующего запуска. Вы хотите на уровне настроек запретить одновременно запускать больше одного экземпляра тайного пайплайна (то есть не меняя настройки остальных) как этого добиться? | A. установить concurrency=1 и parallism=1 на уровне конфига  B.задать всем таскам внутри DAG параметр exclusive\_run=True  C.установить в конструкторе DAG параметр single\_execution=True  ***D.установить в конструкторе DAG параметр max\_active\_runs=1*** |
| У вас определен DAG с schedule\_inretval=’30 12,7\*\*3-5’. Это единственный даг на свежей установкеairflow с настройками по умолчанию. На календаре начало декабря, на часах воскресный полдень, вы находитесь в Москве. Когда будет следующий запуск этого дага по плану? | ***A.12:30 по UTC в понедельник***  B. 12:30 по UTC в среду  C. 12:30 по Москве в среду  D. 17:30 по UTC в среду |
| У вас определен DAG с schedule\_inretval=’30 12,7\*\*3-5’. Это единственный даг на свежей установкеairflow с настройками по умолчанию. На календаре начало декабря, на часах воскресный полдень, вы находитесь в Москве. Когда будет следующий запуск этого дага по плану? | A.12:30 по UTC в понедельник  B. 12:30 по UTC в среду  C. 12:30 по Москве в среду  ***D. 17:30 по UTC в среду*** |
| В Kafka настроен топик с replication factor 6. В двтвцентр ударила молния и две ноды вышли из строя. С соседнего отдела к вам прибежали дата –сайнтисты, которые слышали гром и беспокоятся о кластере и потерянных сообщениях. Что вы им скажете? | A. надо остановить отправку сообщений пока датацентр не поднимет обе ноды  B.они могут слать сообщения так как у нас настроен локальный бэкап на DVD дисках  ***C.пока гром не выведет из строя еще 3 сервера, они могут слать сообщения***  D.надо притормозить пересылку сообщений пока датацентр не поднимет хотя бы одну из двух |
| Ваша кафка-кластер состоит из 10 машин. Вы используете 6 топиков, каждый топик содержит 8 партиций. Сколько всего запущенных процессов брокера будут обрабатывать сообщений? | A. 10 процессов, по одному на каждую ноду  B.80 процессов, по одному процессу на каждую партицию на каждой машне  C.60 процессов, по одному процессу на каждый топик на каждой машине  ***D. 480 процессов по одному процессу на каждый топик-партицию на каждой машине*** |
| Вы работаете в интернет-магазине и провели нагрузочное тестирование вашего кластера кафки из 5 нод. Получилось, что вы можете обрабатывать по 10 миллионов сообщений в минуту. Пообщавшись с маркетингом вы понимаете, что в черную пятницу количество сообщений в очереди вырастет до 25 силлионов в минуту. Перед днем скидок в компании будет code freeze и структура сообщений не изменится. Сколько стоит добавить нод в ваш кластер? | A.25  B.15  C.13  D.5  ***Е.8*** |
| Для каких кейсов Click-House не подходит: | ***A.хранение документов***  B.key-value хранилище  ***C.OLAP***  D.OLTP |
| Выберите верное про Click-House | ***A. хорошо подходит для аналитических запросов***  B.не поддерживает кластерный режим  C.MPP  ***D.колоночная база данных*** |
| Выберите верное про Click-House и репликацию | A. Click-House использует YARN для репликации  B. Click-House занимается репликацией самостоятельно  C.репликация не поддерживается  ***D. Click-House использует кластер Zookeeper для репликации*** |
| Допустим , пользователь , имеющий доступ к системному Docker-демону, запускает контейнер с мотивированием внутрь него корня файловой системы хоста. Какие права на эту самую файловую систему он будет иметь внутри контейнера? | ***A.так нельзя сделать , будет ошибка о некорректном volume***  B.на чтение и на запись  C.только на чтение  D.так нельзя сделать, будет ошибка о недостатке прав |
| Предположим в вашей инфраструктуре все сервисы обернуты в версионированные Docker\_контейнеры. Можно ли что то сделать, чтобы минимизировать нагрузку на сеть в случае развертывания новой версии? | A. да, можно расположить редко меняющиеся и тяжелые операции к началу Dockerfile  ***B. да, можно расположить редко меняющиеся и тяжелые операции к концу Dockerfile***  C. да, можно по максимуму объединить все операции внутри Dockerfile в один слой  D.нет, к сожалению, контейнер слишком атомарен, нагрузку уменьшить не получиться |
| Какая команда Linux позволяет “подменить” корень файловой системы для группы процессов | A. fdisk  ***B.chroot***  C.In-s//other/root  D.cgroups |
| Что из перечисленного относится к ключевым объектам архитектуры Kubernetes? | A. scheduler  B.services  ***C.pods***  ***D.volumes*** |
| Где хранится состояние мастера Kubernetes? | ***A.Kube-apiserver***  B.Kubelet  C.Heapster  D.Etcd |
| Для чего используется Heapster в Kubernetes? | A. инструмент , позволяющий запустить Kubernetes локально на одной ноде  B.платформа для работы с Kubernetes через CLI-интерфейс  ***C.для сбора данных и событий из контейнеров внутри кластера***  D. для создания прокси на кластере |
| Logstash поддерживает разные встроенные плагины для обработки сообщений. Выберите подходящее на ваш взгляд описание для фильтра grok | ***A. Grok позволяет обрабатывать неструктурированную текстовую информацию с помощью определенных пользователем шаблонов***  B.с помощью grok можно мутировать входящие json-докуметы и менять их структуру, так как он делает сам json-десериализацию  C.фильтр grok используется для отправки разных сообщений в соответствующие индексы  D. фильтр grok используется для определения страны входящего сообщения с помощью указания поля, содержащего ID клиента |
| ElasticSearch обещает Near Real Time поиск. Что это значит на практике? | A.каждый новый узел кластера уменьшает скорость ответа уже загруженных данных по функции 1/N  B.данные попадают в индекс за гарантированно небольшую задержку (несколько минут)  C. после индексации становятся доступны для запросов через пару секунд  ***D.за счет распределенного алгоритма индексы пересчитываются за десятки секунд, тогда как в реляционных базах данных это обычно происходит за десятки минут*** |
| Репликации традиционно используют для повышения отказоустойчивости хранения данных в БД. Для каких объектов elasticsearch поддерживает репликации? | A. только индексы , но не конкретные шарды  B.реплицируется только схема документов и типы данных  C.репликация индексов, на деле приводящая к репликам шардов  ***D.каждый документ хранится в отдельном дополнительном индексе с*** другой версией с уникальным UUID в рамках кластера |
| Как с помощью system посмотреть логи Nginx с 9 утра по настоящее время сегоднящнего дня | ***A. journalctl –u nginx.service--since 09:00--until now***  B. systemctl journal nginx.service--start 09:00--end now  C. journalctl –u nginx.service--from 09:00--to now  D. systemctl logs nginx.service--from 09:00--to now |
| В чем отличие Push –модели от Pull-модели при сборе инфраструктурных и бизнес –метрик? | A. в случае Push агенты не нужны вообще, центральный демон опрашивает сервера сам, а в случае Pull он собирает информацию с агентов  ***B. в случае Push агенты пересылают информацию на центральный сервер, в случае Pull центральный сервер собирает информацию с агентов***  C. в случае Push центральный сервер собирает данные с агентов в цикле, в случае Pull – по мере их генерации  D. в случае Push агенты пересылают информацию друг другу, случае Pull центральному серверу напрямую |
| Почему агенты систем мониторинга обычно не рекомендуется запускать в Docker-контейнерах? | A***. потому что это требует пробрасывания в контейнер виртуальных файловых систем хоста, что является плохой практикой и чревато инфраструктурными проблемами***  B. потому что они представляют собой тип процесса, который часто невозможно обернуть в контейнер  C.потому что они не смогут оттуда связаться с главным мониторинг-хостом  D.потому что это потребует серьезного повышения прав у docker -демона |
| Назовите 3 основных оси области видимости задачи SBT? | A. subproject/configurarion/task  B.groupid/artifactid/version  C.local/central/remote  ***D.compile/test/run*** |
| С помощью какого условия блока Post Post в Jenkins можно выполнить код после завершения всех других Post-шагов, независимо от статуса выполнения всего Piopeline: | A.unstable  B.regression  C.cleanup  ***D.always*** |
| Используя какой из приведенных тегов, можно указать Maven. Что данная зависимость будет предоставлена из JDK или других библиотек во время выполнения кода | ***A. <provided>***  B.<exclusion>  C.<optional>  D.<scope> |

**Тестирование Data Engineers (1 вариант)**

|  |  |
| --- | --- |
| Уточните , ML-алгоритмы – ваша специализация? | ***A. да , это моя специализация***  B. нет, это не моя специализация |

**Раздел 27.01 ML-алгоритмы**

|  |  |
| --- | --- |
| Какая из перечисленных метрик характеризует долю правильно предсказанных «единиц» (не «нулей») в задаче бинарной классификации? | A. F1- мера  B.Precision  C.ROC AUC  ***D.Recall***  E. Accuracy |
| Увеличение каких параметров обучения Random Forest может привести к переобучению | A. минимально допустимое количество примеров обучающей выборки в листе дерева  ***B.максимально допустимая глубина дерева***  C. learning rate  D. количество деревьев в ансамбле |
| Какие из перечисленных алгоритмов машинного обучения решают задачу классификации | A.PCA  ***B.Random Forest***  C.t-SNE  ***D. градиентный бустинг решающих деревьев***  ***Е. метод ближайших соседей***  F.k-means  ***G.Support Vector Machine*** |

**Раздел 28,01. Библиотеки для Data Science**

|  |  |
| --- | --- |
| Какой тип для sparse-матриц в scipy позволяет эффективно производить выборку по строкам | ***A.scipy.sparse.csr\_matrix***  B. scipy.sparse.ccc\_matrix  C. scipy.sparse.coo\_matrix  D. scipy.sparse.csc\_matrix |
| Какие из вариантов возвращают numpy-матрицу размера 3х3 , заполненную значениями True? | ***A. numpy.ones((3,3), dtype=bool)***  ***B. numpy.full((3,3), True,dtype=bool)***  C. numpy.bool((3,3))  D.~ numpy.zeros(3, dtype=bool) |
| Какие из перечисленных библиотек для python являются фреймворками для обучения глубоких нейронных сетей? | ***A. pytorch***  B.pandas  C.scikit-learn  D.scipy  ***E.theano***  F.pyspark  ***G. tensorflow*** |

**Раздел 29.01. Практики CI/CD**

|  |  |
| --- | --- |
| В чем может проявиться преимущество монорепозитория для больших распределенных команд | A. так, как правило, будет гораздо проще проводить code review  B.проект сильно быстро собирается  ***C.все проекты используют одну и ту же версию основных shared-зависимостей – последнюю***  D. надо склонировать код всего лишь один раз |
| В чем основная задача системы управления конфигурацией? | A. проверка доступности различных компонентов инфраструктуры  B.установка пакетов и создание пользователей  C.приведение системы к состоянию, декларативно описанному в рецептах  ***D.создание конфигурационных файлов для сервисов*** |
| Какой механизм в дополнение к автоматическим тестам при мердже часто используется для проверки работоспособности стабильных бранчей проекта? | ***A. расширенные интеграционные тесты***  B. ручное тестирование  C.smoke-тесты  D. «ночные» сборки |

**Раздел 30.01. DevOps**

|  |  |
| --- | --- |
| Проект начинающийся с нуля и позволяющий с самого начала использовать желаемый набор практик, часто называют термином, пришедшим в IT из градостроительства – green field (зеленое поле, беспроблемный участок для действий) А как называют legacy-площадку, скрывающую возможные проблемы? | A.white field  B.blue field  C. black field  ***D. brown field*** |
| Существует одна из вариаций Закона Конвея: «если у вас четыре независимые команды работают над компилятором, то вы получите четырехпроходной компилятор» Какую типичную проблему процесса разработки описывает этот закон? | A.недостаток «вертикального» взаимодействия между людьми (больше проблем должны эскалироваться наверх вместо решения на местах)  B.неизбежность накопления технического долга  ***C.большие проекты неизбежно переходят на микросервисную архитектуру***  D.недостаток горизонтального взаимодействия между людьми (больше проблем должны решаться на местах вместо эскалирования их наверх |
| Что такое технический долг? | A. не доделанная до конца работа по внесению изменений (правка документации, тестирование ит.д)  B.набор фич продукта, которые присутствуют в роадмапе, но не внесены в текущий релиз  ***C.совокупность всех незакрытых тикеров в текущем релизе на данный момент***  D. количество денег , которое компания потеряла в результате неэффективных ITпроцессов |