УТВЕРЖДАЮ

Зав. кафедрой СП ЮУрГУ

\_\_\_\_\_\_\_\_\_Л.Б. Соколинский

«\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_2022 г.

Фонд оценочных средств

ООП «Инженерия информационных и интеллектуальных систем»

по направлению 09.03.04 – Программная инженерия

Дисциплина «Основы распределенной обработки данных»

| **№ КМ** | **Вид КМ** | **Наименование КМ** | **Оценочные средства** |
| --- | --- | --- | --- |
|  | Текущий контроль | Минитест 1: большие данные | Вопросы для подготовки к тесту:  1. Укажите определение больших данных от компании Gartner.  2. Отметьте способы обработки больших данных.  3. Какой из примеров можно назвать большими данными?  4. Производительность какого вида компьютерной памяти растет медленнее других?  5. Отметьте, каким из указанных программным продуктом можно обрабатывать большие данные. |
|  | Текущий контроль | Минитест 2: Hadoop | Вопросы для подготовки к тесту:  1. Укажите разработчика Hadoop.  2. Укажите системные принципы Hadoop.  3. Какие функции выполняют вендоры дистрибутивов Hadoop?  4. Что такое горизонтальное/вертикальное масштабирование?  5. На основе идей какой системы был разработан Hadoop? |
|  | Текущий контроль | Минитест 3: HDFS | Вопросы для подготовки к тесту:  1. На основе идей какой системы была разработана HDFS?  2. Укажите отличия файловой системы и распределенной файловой системы.  3. Укажите, для каких задач подходит HDFS?  4. Какие фоновые процессы работают в HDFS?  5. Вычислите, каким должен быть размер блока данных в HDFS, чтобы seek time, равный 10 ms, достигал всего 1% от transfer rate, равного 100Mb/s. |
|  | Текущий контроль | Минитест 4: MapReduce | Вопросы для подготовки к тесту:  1. Как определяется число mapper-процессов в MapReduce задаче?  2. Где хранятся промежуточные данные между шагами Map и Reduce?  3. Как называется мастер-процесс в MapReduce Hadoop?  4. Какие опциональные функции может реализовать разработчик задачи MapReduce в Hadoop?  5. Укажите, какими недостатками обладают MapReduce алгоритмы на графах. |
|  | Текущий контроль | Минитест 5: Pig и Hive | Вопросы для подготовки к тесту:  1. Из каких компонент состоит Pig?  2. Какой продукт в Hadoop включает в себя реляционную СУБД для хранения метаданных?  3. Напишите команду запуска с помощью Pig скрипта test.pig из терминала в локальном режиме.  4. Выберите элементы модели данных Hive.  5. Как называется команда в Pig, которая выводит историю (шаги) создания указанного bag? |
|  | Текущий контроль | Минитест 6: HBase | Вопросы для подготовки к тесту:  1. Что хранится в HFile СУБД HBase?  2. Напишите запрос для СУБД HBase для создания таблицы Writers с семействами столбцов author и books.  3. Какая команда в СУБД HBase вставляет новые данные в таблицу?  4. Какой тип должен быть у ключа в таблице HBase?  5. Какие функции выполняет HMaster в СУБД HBase ? |
|  | Текущий контроль | Минитест 7: YARN | Вопросы для подготовки к тесту:  1. Какие недостатки есть у фреймворка MapReduce v1?  2. Как расшифровывается YARN?  3. Как называется демон в YARN, отвечающий за ресурсы на каждом сервере кластера?  4. Какие функции выполняет Node Manager?  5. Укажите те функции JobTracker, которые перешли к MRAppMaster. |
|  | Текущий контроль | Минитест 8: Apache Kafka, Zookeeper | Вопросы для подготовки к тесту:  1. Какие функции выполняет Apache Kafka?  2. Какие функции у брокера Apache Kafka?  3. Что такое topic в Apache Kafka?  4. Укажите какой компонент Apache Kafka взаимодействует с Zookeeper.  5. Какую задачу выполняет Request processor в ZooKeeper? |
|  | Текущий контроль | Минитест 9: Apache Spark | Вопросы для подготовки к тесту:  1. Какими преимуществами обладает Spark по сравнению с MapReduce?  2. Как в Spark называется абстрактное представление распределенной RAM?  3. Отметьте операторы actions в приведенном списке  4. Как называется тип переменных в Spark, с помощью которых можно реализовать счетчики в MapReduce?  5. Что будет делать Spark, если в памяти недостаточно места для новых партиций RDD? |
|  | Текущий контроль | Минитест 10: анализ и визуализация данных | Вопросы для подготовки к тесту:  1. Как в MLlib называются cтолбцы в датасете?  2. Если в датасете по некоторому свойству 90% значений от 0 до 1, а остальные 10% - от 1 до 1 000 000, то для корректности решения задачи классификации необходимо провести одно из следующих действий: нормализация, трансформация, кластеризация. Укажите какое.  3. Укажите, что необходимо сделать перед подачей датасета в модель MLlib?  =перевести текстовые значения в числовые  =перевести числовые значения в текстовые  =переставить местами строки и столбцы в датасете  4. Укажите иерархию между объектами рисунка Figure, Axes, Axis в matplotlib.  5. Как отобразить легенду к графику в matplotlib? |
|  | Текущий контроль | ПЗ 1. Установка Hadoop. Работа с HDFS | Вопросы для подготовки к устному опросу:   1. Какие функции выполняет NameNode в Hadoop? 2. Что такое фактор репликации? 3. На сколько блоков разбиваются данные при загрузке в HDFS? 4. Какими командами можно считать данные из HDFS? |
|  | Текущий контроль | ПЗ 2. Разработка MapReduce-приложения | Вопросы для подготовки к устному опросу:   1. Какие функции выполняет ResourceManager в Hadoop? 2. Из каких частей состоит приложение MapReduce? 3. Что такое shuffle? |
|  | Текущий контроль | ПЗ 3. Разработка статистических отчетов с использованием Apache Hive | Вопросы для подготовки к устному опросу:   1. Какой язык запросов используется в Hive? 2. Что хранится в Metastore? 3. Может ли Hive хранить данные в HDFS? |
|  | Текущий контроль | ПЗ 4. Разработка приложения в Apache Spark | Вопросы для подготовки к устному опросу:   1. Что такое RDD в Spark? 2. Что в Spark означает термин «ленивые вычисления»? 3. Что выполняет метод parallelize в Spark? |
|  | Текущий контроль | ПЗ 5. Разработка приложения для анализа и визуализации данных в Hadoop | Вопросы для подготовки к устному опросу:   1. Укажите иерархию между объектами рисунка Figure, Axes, Axis в matplotlib. 2. Как отобразить легенду к графику в matplotlib? 3. Какая команда в matplotlib рисует круговую диаграмму? |
|  | Промежуточный аттестация | Итоговое тестирование | Вопросы для подготовки к экзамену:   1. Определение больших данных. Проблема больших данных. 2. Система хранения больших данных: архитектура системы хранения (DAS, SAN, NAS), структура хранения (блочное, объектное, облачное, файловая система), реализация системы хранения (SQL, NoSQL, newSQL, хранилища и озера данных), устройства хранения (кеш-память, оперативная память, HDD/SSD). 3. Платформа Hadoop. Компоненты Hadoop. Экосистема Hadoop. Дистрибутивы Hadoop. 4. Архитектура HDFS. 5. Hadoop MapReduce. 6. Обработка данных в Hadoop с помощью Pig и СУБД Apache Hive. 7. СУБД HBase. NoSQL-системы. 8. Обработка больших данных в Apache Spark. 9. Hadoop YARN. HDFS Federation. HDFS High Availability. Архитектура YARN. 10. ZooKeeper. 11. Управление потоками данных в Hadoop: Apache Kafka. 12. Разработка собственных приложений для управления и анализа данных в Hadoop. Визуализация данных. |

Паспорт фонда оценочных средств приведен в п. 6.3 РПД.

Разработчик Е.В. Иванова

ФГАОУ ВО «Южно-Уральский государственный университет   
(национальный исследовательский университет)»

Кафедра системного программирования

Дисциплина «Основы распределенной обработки данных»

ИТОГОВЫЙ ТЕСТ

| № | Вопрос | Варианты ответа |
| --- | --- | --- |
|  | Какие функции выполняют дистрибутивы Hadoop? | |  |  | | --- | --- | |  | Предоставляют Hadoop в облаке | |  | Представляют собой аппаратно-программные комплексы | |  | Обеспечивают совместимость разных компонент Hadoop | |  | Содержит компоненты Hadoop с исправленными ошибками | |
|  | Как осуществляется чтение данных из HDFS? | |  |  | | --- | --- | |  | последовательное чтение всего файла | |  | произвольный доступ к нужной части файла | |
|  | Для какого шаблона взаимодействия с данными был разработан HDFS? | |  |  | | --- | --- | |  | один раз записали - один раз прочитали | |  | один раз записали - много раз прочитали | |  | много раз изменили данные - много раз прочитали | |  | много раз изменили данные - один раз прочитали | |
|  | Задача. Пусть имеется два процесса Client 1 и Client 2, которые обновляют переменную X в реплицируемой базе данных. Выполняются следующие операции:  Время Процесс Операция  09:20 Client 1 X=A  09:25 Client 1 X=B  09:30 Client 2 X=C  09:35 Client 1 считать X  Какое(-ие) значение(-я) X может получить Client 1 при согласованности read-your-writes в БД? | |  |  | | --- | --- | |  | A | |  | B | |  | C | |  | (пусто) | |
|  | Сколько всего блоков будет храниться в HDFS для файла размером 2400 Мб, если размер блока 128 Мб, число серверов 10 шт., фактор репликации 2? | |  |  | | --- | --- | |  | 18 | |  | 19 | |  | 20 | |  | 36 | |  | 38 | |  | 180 | |  | 190 | |
|  | Какие элементы можно встретить в DAG в Spark? | |  |  | | --- | --- | |  | RegionServer | |  | Map | |  | data sink | |  | Операторы | |  | DataNode | |
|  | Укажите три верных утверждения про HBase. | |  |  | | --- | --- | |  | может хранить несколько версий для данных | |  | является реляционной СУБД | |  | работает только на одном узле Hadoop-кластера | |  | работает поверх HDFS | |  | записи старше заданного промежутка времени автоматически удаляются | |
|  | Укажите две особенности MapReduce v2.0. | |  |  | | --- | --- | |  | HDFS Federation | |  | единственный сервер NameNode | |  | поддержка мультиарендности | |  | единственный процесс JobTracker | |
|  | Укажите Python-функцию для построения круговой диаграммы. | |  |  | | --- | --- | |  | pie | |  | xlabel | |  | plot | |  | bar | |
|  | Отметьте две функции, которые выполняет HMaster. | |  |  | | --- | --- | |  | Отдает клиенту данные из таблицы по ключу | |  | Возвращает клиенту имя региона и RegionServer по ключу | |  | Пишет в HLog изменения в таблице | |  | Перераспределяет регионы по серверам | |
|  | Как называется свойство в теореме CAP, обозначающее возможность сбоя в системе? | |  |  | | --- | --- | |  | доступность | |  | согласованность | |  | устойчивость к разделению | |
|  | Что будет делать Spark, если в памяти недостаточно места для новых партиций RDD? | |  |  | | --- | --- | |  | Выдаст ошибку и не запишет новые данные | |  | Вытеснит из памяти часто используемые партиции, чтобы хватило место для новых | |  | Вытеснит из памяти редко используемые партиции, чтобы хватило место для новых | |  | Удалит из памяти наиболее близкие по размеру партиции | |
|  | Отметьте верное утверждение про DataNode. | |  |  | | --- | --- | |  | DataNode хранит дерево файлов | |  | DataNode хранит блоки данных | |  | DataNode хранит реплики данных | |  | DataNode - это один и только один узел в кластере | |
|  | Как определяется количество task'ов (т.е. процессов map и reduce) в Hadoop MapReduce ver.1? | |  |  | | --- | --- | |  | определяется динамически во время запуска задачи | |  | прописывается в конфигурационном файле, для изменения требуется перезапуск MapReduce | |
|  | Какую задачу выполняет Request processor в ZooKeeper? | |  |  | | --- | --- | |  | обрабатывает запросы на запись | |  | хранит дерево данных ZooKeeper | |  | реализует atomic broadcast | |
|  | Какая команда создает в Hive таблицу users с полями имя (name), эл. почта (email), возраст (age) и использует в качестве разделителя ';', а в качестве хранилища текстовый файл. | |  |  | | --- | --- | |  | create table users (name string, email string, age int) row format delimited fields terminated by ';' stored as textfile; | |  | create table users (fio string, email string, age int) row format delimited fields terminated by ',' stored as textfile; | |  | create table users (name string, email string, age int) delimited fields terminated by ';'; | |  | create table users (name string, email string, age int) row format delimited fields terminated ';' stored as textfile; | |
|  | Даны следующие настройки Kafka:  replication-factor = 5  min.insync.replicas = 3  Сколько НЕсинхронных реплик содержится в системе? | |  |  | | --- | --- | |  | 0 | |  | 2 | |  | 3 | |  | 5 | |
|  | Как называется команда в Pig, которая выводит историю (шаги) создания указанного Bag? | |  |  | | --- | --- | |  | dump | |  | illustrate | |  | describe | |  | store | |  | load | |
|  | Отметьте продукты экосистемы Hadoop, которые работают "поверх" MapReduce, т.е. предоставляют простой интерфейс для написания MapReduce-приложений. | |  |  | | --- | --- | |  | HBase | |  | Pig | |  | HDFS | |  | Hive | |
|  | Может ли в HDFS Federation узел DataNode хранить блоки из разных пространств имен? | |  |  | | --- | --- | |  | да | |  | нет | |