Приложение 5

Федеральное государственное бюджетное образовательное учреждение высшего образования «Тамбовский государственный университет имени Г.Р. Державина»
Институт экономики, управления и сервиса
Кафедра математического моделирования и информационных технологий

УТВЕРЖДАЮ:
Директор института

Е. Ю. Меркулова
«23» июня 2023 г.

РАБОЧАЯ ПРОГРАММА

по дисциплине Б1.В.ДВ.01.1 Анализ больших данных

Направление подготовки/специальность: 38.04.01 - Экономика

Профиль/направленность/специализация: Стратегический бизнес-анализ и аудит в цифровой экономике

Уровень высшего образования: магистратура

Квалификация: Магистр

год набора: 2023

Тамбов, 2023

Автор программы:
Кандидат педагогических наук, доцент Клыгина Елена Владимировна

Рабочая программа составлена в соответствии с ФГОС ВО по направлению подготовки 38.04.01 - Экономика (уровень магистратуры) (приказ Министерства науки и высшего образования РФ от «11» августа 2020 г. № 939).

Рабочая программа принята на заседании Кафедры математического моделирования и информационных технологий «16» июня 2023 г. Протокол № 15

Рассмотрена и одобрена на заседании Ученого совета Института экономики, управления и сервиса, Протокол от «23» июня 2023 г. № 12.

СОДЕРЖАНИЕ

1. Цели и задачи дисциплины…… 4

2. Место дисциплины в структуре ОП Магистратуры……... 5

3. Объем и содержание дисциплины…… 5

4. Контроль знаний обучающихся и типовые оценочные средства…… 8

5. Методические указания для обучающихся по освоению дисциплины (модуля)…….. 15

6. Учебно-методическое и информационное обеспечение дисциплины…… 17

7. Материально-техническое обеспечение дисциплины, программное обеспечение, профессиональные базы данных и информационные справочные системы……. 18

1. Цели и задачи дисциплины

1.1 Цель дисциплины – формирование компетенций:

ПК-2 Способность проводить мониторинг и анализировать бизнес-процессы на основе цифровых технологий, оценивать возможности организации, необходимые для проведения стратегических изменений

1.2 Типы задач профессиональной деятельности, к которым готовятся обучающиеся в рамках освоения дисциплины:

- аналитический

1.3 Дисциплина ориентирована на подготовку обучающихся к профессиональной деятельности в сфере: 08 Финансы и экономика (в сферах: исследований, анализа и прогнозирования социально-экономических процессов и явлений на микроуровне и макроуровне в экспертно-аналитических службах (центрах экономического анализа, правительственном секторе, общественных организациях); производства продукции и услуг, включая анализ спроса на продукцию и услуги, и оценку их текущего и перспективного предложения, продвижение продукции и услуг на рынок, планирование и обслуживание финансовых потоков, связанных с производственной деятельностью; кредитования; страхования, включая пенсионное и социальное; операций на финансовых рынках, включая управление финансовыми рисками; внутреннего и внешнего финансового контроля и аудита, финансового консультирования; консалтинга)

1.4 В результате освоения дисциплины у обучающихся должны быть сформированы:

Обобщенные трудовые функции / трудовые функции / трудовые или профессиональные действия (при наличии профстандарта)	Код и наименование компетенции ФГОС ВО, необходимой для формирования трудового или профессионального действия	Индикаторы достижения компетенций
	ПК-2 Способность проводить мониторинг и анализировать бизнес-процессы на основе цифровых технологий, оценивать возможности организации, необходимые для проведения стратегических изменений	Анализирует большие данные и использует результаты анализа в процессе проведения стратегических изменений в организации

1.5 Согласование междисциплинарных связей дисциплин, обеспечивающих освоение компетенций:

ПК-2 Способность проводить мониторинг и анализировать бизнес-процессы на основе цифровых технологий, оценивать возможности организации, необходимые для проведения стратегических изменений

№ п/п	Наименование дисциплин, определяющих междисциплинарные связи	Форма обучения		
		Очная (семестр)		
		2	3	4
1	Анализ и оценка бизнес-процессов			+
2	Анализ финансовой отчетности			+
3	Базы данных	+		

4	Инвестиционный анализ		+	
5	Ознакомительная практика	+		
6	Практика по профилю профессиональной деятельности		+	+
7	Управление ресурсами			+
8	Финансовый анализ		+	
9	Цифровая экономика	+		
10	Цифровые технологии управления бизнесом	+		

2. Место дисциплины в структуре ОП магистратуры:

Дисциплина «Анализ больших данных» относится к части, формируемой участниками образовательных отношений, учебного плана ОП по направлению подготовки 38.04.01 - Экономика.

Дисциплина «Анализ больших данных» изучается в 2 семестре.

3.Объем и содержание дисциплины

3.1.Объем дисциплины: 2 з.е.

Очная: 2 з.е.

Вид учебной работы	Очная (всего часов)
Общая трудоёмкость дисциплины	**72**
Контактная работа	32
Лекции (Лекции)	16
Практические (Практ. раб.)	16
Самостоятельная работа (СР)	40
Зачет	-

3.2.Содержание курса:

№ темы	Название раздела/темы	Вид учебной работы, час.			Формы текущего контроля
		Лекции	Практ. раб.	СР	
		О	О	О	
2 семестр					
1	Модель программирования Hadoop.	2	2	5	Лабораторная работа
2	Распределенная обработка данных с помощью Hadoop.	1	1	5	Лабораторная работа

3	Обработка данных с помощью модуля Hadoop Common.	1	1	5	Лабораторная работа
4	Обработка данных с помощью модуля Hadoop YARN.	2	2	4	Лабораторная работа
5	Обработка данных с помощью модуля Hadoop MapReduce.	2	2	4	Лабораторная работа
6	Тестирование файловой системы HDFS.	2	2	4	Лабораторная работа
7	Архитектура Spark. Хранилище данных	2	2	4	Лабораторная работа
8	Архитектура Spark. API.	2	2	5	Лабораторная работа
9	Архитектура Spark. Менеджер кластера.	2	2	4	Лабораторная работа; Тестирование

Тема 1. Модель программирования Hadoop. (ПК-2)

Лекция.

Hadoop – Обзор больших данных. Что такое большие данные. Что входит в большие данные. Преимущества больших данных.

Практическое занятие.

1. Основы Apache Hadoop. Развертывание и тестирование.

Задания для самостоятельной работы.

Задачи:

1. Подготовка к защите лабораторной работы.

Тема 2. Распределенная обработка данных с помощью Hadoop. (ПК-2)

Лекция.

Первоначальная настройка. Запуск Hadoop. Проверка файловой системы HDFS. Тестирование Hadoop. Web-интерфейсы.

Практическое занятие.

1. Создать программу для построения инвертированного индекса по набору текстовых документов.

Задания для самостоятельной работы.

Задачи:

1. Подготовка к защите лабораторной работы.

Тема 3. Обработка данных с помощью модуля Hadoop Common. (ПК-2)

Лекция.

Hadoop Common — набор библиотек, сценариев и утилит для создания инфраструктуры, аналог командной строки.

Практическое занятие.

1. Создать программу подсчета частоты встречаемости слов в тексте.

Задания для самостоятельной работы.

Задачи:

1. Подготовка к защите лабораторной работы.

Тема 4. Обработка данных с помощью модуля Hadoop YARN. (ПК-2)

Лекция.

Что такое YARN, для чего он нужен. YARN и старый MapReduce. Компоненты MapReduce на YARN. Выполнение MR-задачи на YARN.

Практическое занятие.

1. Создать программу с реализацией функций Map и Reduce в платформе Apache Hadoop.

2. Создать программу для вычисления по каждому пользователю общего количества посещенных ими сайтов.

Задания для самостоятельной работы.

Задачи:

1. Подготовка к защите лабораторной работы.

Тема 5. Обработка данных с помощью модуля Hadoop MapReduce. (ПК-2)

Лекция.

Обработка больших данных: первые шаги в понимании Hadoop MapReduce. Hadoop MapReduce и что его окружает.

Практическое занятие.

1. Создать программу поиска кратчайших путей в графе с использованием MapReduce.

Задания для самостоятельной работы.

Задачи:

1. Подготовка к защите лабораторной работы.

Тема 6. Тестирование файловой системы HDFS. (ПК-2)

Лекция.

Задачи, для которых подходит и не подходит HDFS. Демоны HDFS. Файлы и блоки. Репликация блоков. Клиенты, Namenode и Datanodes. Чтение и запись файла. Namenode: использование памяти. Устойчивость к отказам в Namenode. Доступ к HDFS, в том числе через прокси. Команды оболочки shell. Копирование данных в shell, удаление и статистика. Команда fsck. Права в HDFS. Команда DFSAdmin. Балансер. File System Java API. Реализация File System. Объект Configuration. Чтение данных из файла и запись в него. Подстановки (globbing).

Практическое занятие.

1. Сконфигурировать HDFS, установив количество копий в 4 (репликация), и проверить работу файловой системы с этой конфигурацией.

Задания для самостоятельной работы.

Задачи:

1. Подготовка к защите лабораторной работы.

Тема 7. Архитектура Spark. Хранилище данных (ПК-2)

Лекция.

Как работает распределенная среда Spark: основные особенности архитектуры. Драйвер Spark. Spark-исполнители.

Практическое занятие.

Написать программы для Spark, решающие следующие задачи:

1. На основе заданного текста в файле составить словарь — то есть список всех уникальных слов, используемых в тексте в алфавитном порядке. Знаки препинания игнорировать.

2. На основе заданного текста подсчитать вхождение каждого слова. То есть словарь в котором для каждого слова указывается сколько раз оно входит в текст.

Задания для самостоятельной работы.

Задачи:

1. Подготовка к защите лабораторной работы.

Тема 8. Архитектура Spark. API. (ПК-2)

Лекция.

Рабочий процесс и характеристики. Искровые рабочие характеристики. Общие термины.

Практическое занятие.

Написать программы для Spark, решающие следующие задачи:

1. Даны два разных текста. Сформировать 3 множества: пересечение словарей заданных текстов, множество слов, входящих только в первый текст, множество слов, входящих только во второй текст.

Задания для самостоятельной работы.

Задачи:

1. Подготовка к защите лабораторной работы.

Тема 9. Архитектура Spark. Менеджер кластера. (ПК-2)

Лекция.

Архитектура Spark кластера. Искровая операционная архитектура. Режим управления кластером. Мониторинг. Планирование работы и задач. Связанные термины.

Практическое занятие.

Написать программы для Spark, решающие следующие задачи:

1. Дана матрица чисел. Проверить является ли матрица симметричной относительно главной диагонали.
2. Написать программу перемножения двух матриц. .

Задания для самостоятельной работы.

Задачи:

1. Подготовка к защите лабораторной работы.

4. Контроль знаний обучающихся и типовые оценочные средства

4.1. Распределение баллов:

2 семестр

- текущий контроль – 80 баллов
- контрольные срезы – 2 среза по 10 баллов каждый
- премиальные баллы – 20 баллов

Распределение баллов по заданиям:

№ темы	Название темы / вид учебной работы	Формы текущего контроля / срезы	Max. кол-во баллов	Методика проведения занятия и оценки
1.	Модель программирования Hadoop.	Лабораторная работа	10	Выполнение и защита лабораторной работы. 9-10 баллов – студент выполнил работу без ошибок и недочетов, допустил не более одного недочета. 7-8 баллов – студент выполнил не менее половины работы. 5-6 балла – студент правильно выполнил не менее половины работы, но не продвинулся в решении задачи. 3-4 балла– студент правильно выполнил более 25%, но менее 50% работы, имеет продвижение в решении задач. 1-2 балл – студент правильно выполнил менее 25% работы.

2.	Распределенная обработка данных с помощью Hadoop.	Лабораторная работа	10	Выполнение и защита лабораторной работы. 9-10 баллов – студент выполнил работу без ошибок и недочетов, допустил не более одного недочета. 7-8 баллов – студент выполнил не менее половины работы. 5-6 балла – студент правильно выполнил не менее половины работы, но не продвинулся в решении задачи. 3-4 балла– студент правильно выполнил более 25%, но менее 50% работы, имеет продвижение в решении задач. 1-2 балл – студент правильно выполнил менее 25% работы.
3.	Обработка данных с помощью модуля Hadoop Common.	Лабораторная работа	10	Выполнение и защита лабораторной работы. 9-10 баллов – студент выполнил работу без ошибок и недочетов, допустил не более одного недочета. 7-8 баллов – студент выполнил не менее половины работы. 5-6 балла – студент правильно выполнил не менее половины работы, но не продвинулся в решении задачи. 3-4 балла– студент правильно выполнил более 25%, но менее 50% работы, имеет продвижение в решении задач. 1-2 балл – студент правильно выполнил менее 25% работы.
4.	Обработка данных с помощью модуля Hadoop YARN.	**Лабораторная работа(контрольный срез)**	10	Выполнение и защита лабораторной работы. В случае успешного выполнения всех заданий лабораторной работы студент получает 10 баллов
5.	Обработка данных с помощью модуля Hadoop MapReduce.	Лабораторная работа	10	Выполнение и защита лабораторной работы. 9-10 баллов – студент выполнил работу без ошибок и недочетов, допустил не более одного недочета. 7-8 баллов – студент выполнил не менее половины работы. 5-6 балла – студент правильно выполнил не менее половины работы, но не продвинулся в решении задачи. 3-4 балла– студент правильно выполнил более 25%, но менее 50% работы, имеет продвижение в решении задач. 1-2 балл – студент правильно выполнил менее 25% работы.
6.	Тестирование файловой системы HDFS.	Лабораторная работа	10	Выполнение и защита лабораторной работы. 9-10 баллов – студент выполнил работу без ошибок и недочетов, допустил не более одного недочета. 7-8 баллов – студент выполнил не менее половины работы. 5-6 балла – студент правильно выполнил не менее половины работы, но не продвинулся в решении задачи. 3-4 балла– студент правильно выполнил более 25%, но менее 50% работы, имеет продвижение в решении задач. 1-2 балл – студент правильно выполнил менее 25% работы.
7.	Архитектура Spark. Хранилище данных	Лабораторная работа	10	Выполнение и защита лабораторной работы. 9-10 баллов – студент выполнил работу без ошибок и недочетов, допустил не более одного недочета. 7-8 баллов – студент выполнил не менее половины работы. 5-6 балла – студент правильно выполнил не менее половины работы, но не продвинулся в решении задачи. 3-4 балла– студент правильно выполнил более 25%, но менее 50% работы, имеет продвижение в решении задач. 1-2 балл – студент правильно выполнил менее 25% работы.
8.	Архитектура Spark. API.	Лабораторная работа	10	Выполнение и защита лабораторной работы. 9-10 баллов – студент выполнил работу без ошибок и недочетов, допустил не более одного недочета. 7-8 баллов – студент выполнил не менее половины работы. 5-6 балла – студент правильно выполнил не менее половины работы, но не продвинулся в решении задачи. 3-4 балла– студент правильно выполнил более 25%, но менее 50% работы, имеет продвижение в решении задач. 1-2 балл – студент правильно выполнил менее 25% работы.

9.	Архитектура Spark. Менеджер кластера.	Лабораторная работа	10	Выполнение и защита лабораторной работы. 9-10 баллов – студент выполнил работу без ошибок и недочетов, допустил не более одного недочета. 7-8 баллов – студент выполнил не менее половины работы. 5-6 балла – студент правильно выполнил не менее половины работы, но не продвинулся в решении задачи. 3-4 балла– студент правильно выполнил более 25%, но менее 50% работы, имеет продвижение в решении задач. 1-2 балл – студент правильно выполнил менее 25% работы.
		Тестирование(контрольный срез)	10	В случае правильных ответов на 51% заданий тестирования студент получает 10 баллов
10.	Премиальные баллы		20	Дополнительные премиальные баллы могут быть начислены: - постоянная активность во время практических занятий – 10 баллов; - полностью подготовленная к публикации статья по тематике в рамках дисциплины – 10 баллов; - публикация статьи по тематике изучаемой дисциплины в сборнике студенческих работ / материалах международной, всероссийской конференции / журнале из перечня ВАК – 10 / 15 / 20
11.	Индивидуальные задания, с помощью которых можно набрать дополнительные баллы		100	Добор: студент может предоставить все задания текущего контроля и контрольные срезы
12.	Итого за семестр		100	

Итоговая оценка по зачету выставляется в 100-балльной шкале и в традиционной четырехбалльной шкале. Перевод 100-балльной рейтинговой оценки по дисциплине в традиционную четырехбалльную осуществляется следующим образом:

100-балльная система	Традиционная система
50 - 100 баллов	Зачтено
0 - 49 баллов	Не зачтено

4.2 Типовые оценочные средства текущего контроля

Лабораторная работа

Тема 1. Модель программирования Hadoop.

Задания к лабораторной работе по теме №1 "Модель программирования Hadoop".

1 1. Основы Apache Hadoop. Развертывание и тестирование.

Тема 2. Распределенная обработка данных с помощью Hadoop.

Задания к лабораторной работе по теме №2 "Распределенная обработка данных с помощью Hadoop".

1. Создать программу для построения инвертированного индекса по набору текстовых документов.

Тема 3. Обработка данных с помощью модуля Hadoop Common.

Задания к лабораторной работе по теме №3 "Обработка данных с помощью модуля Hadoop Common".

1. Создать программу подсчета частоты встречаемости слов в тексте.

Тема 4. Обработка данных с помощью модуля Hadoop YARN.

Задания к лабораторной работе по теме №4 "Обработка данных с помощью модуля Hadoop YARN".

1. Создать программу с реализацией функций Map и Reduce в платформе Apache Hadoop.

2. Создать программу для вычисления по каждому пользователю общего количества посещенных ими сайтов.

Тема 5. Обработка данных с помощью модуля Hadoop MapReduce.

Задания к лабораторной работе по теме №5 "Обработка данных с помощью модуля Hadoop MapReduce".

1. Создать программу поиска кратчайших путей в графе с использованием MapReduce.

Тема 6. Тестирование файловой системы HDFS.

Задания к лабораторной работе по теме №6 "Тестирование файловой системы HDFS".

1. Сконфигурировать HDFS, установив количество копий в 4 (репликация), и проверить работу файловой системы с этой конфигурацией.

Тема 7. Архитектура Spark. Хранилище данных

Задания к лабораторной работе по теме №7 "Архитектура Spark. Хранилище данных".

Написать программы для Spark, решающие следующие задачи:

1. На основе заданного текста в файле составить словарь — то есть список всех уникальных слов, используемых в тексте в алфавитном порядке. Знаки препинания игнорировать.

2. На основе заданного текста подсчитать вхождение каждого слова. То есть словарь в котором для каждого слова указывается сколько раз оно входит в текст.

Тема 8. Архитектура Spark. API.

Задания к лабораторной работе по теме №8 "Архитектура Spark. API".

Написать программы для Spark, решающие следующие задачи:

1. Даны два разных текста. Сформировать 3 множества: пересечение словарей заданных текстов, множество слов, входящих только в первый текст, множество слов, входящих только во второй текст.

Тема 9. Архитектура Spark. Менеджер кластера.

Задания к лабораторной работе по теме №9 "Архитектура Spark. Менеджер кластера".

Написать программы для Spark, решающие следующие задачи:

1. Дана матрица чисел. Проверить является ли матрица симметричной относительно главной диагонали.

2. Написать программу перемножения двух матриц. .

Тестирование

Тема 9. Архитектура Spark. Менеджер кластера.

Тестовые задания

1. Каковы наиболее общие форматы входных данных, определённые в Hadoop? Какой из них используется по умолчанию?

Ответ: Наиболее общие форматы входных данных, определённые в Hadoop это

- TextInputFormat

- KeyValueInputFormat
- SequenceFileInputFormat

По умолчанию используется TextInputFormat.

2. В чём отличие между классами TextInputFormat и KeyValueInputFormat?

Ответ: TextInputFormat читает строки из текстового файла и предоставляет смещение строки относительно начала файла в качестве входного ключа для отображения (Mapper), а саму строку в качестве входного значения для отображения (Mapper).

KeyValueInputFormat читает строки из текстового файла и производит разбор каждой отдельной строки в пару ключ-значение. Все символы от начала строки и до первого символа табуляции передаются отображению (Mapper) в качестве ключа, а остаток передаётся отображению (Mapper) в качестве значения.

3. Что такое трекер заданий (JobTracker)?

Ответ: Трекер заданий (JobTracker) — это сервис платформы Hadoop, который запускает на кластере задания (Job), построенные согласно модели вычисления MapReduce. Трекером заданий (JobTracker) также называется узел кластера, на котором работает указанный сервис.

4. Что такое Hadoop Streaming?

Ответ: Hadoop Streaming — это общий API, позволяющий программе, написанной теоретически на любом языке, реализовать фазы отображения (Map) и свёртки (Reduce) в платформе Hadoop.

5. Как выполнить проверку работоспособности файловой системы HDFS?

Ответ: Для проверки работоспособности файловой системы HDFS используется утилита FSCK. Она очень удобна для проверки целостности файла, имён блоков и расположения блоков.

Пример. Запуск утилиты FSCK. hdfs fsck /dir/hadoop-test -files -blocks -locations

6. Каковы параметры функций отображений (map) и свёртки (reducer)?

Ответ: Сигнатуры методов отображения и свёртки много говорят о типе входных и выходных данных, которыми работает задание (Job). В предположении что вы используете TextInputFormat, параметры функции отображения (Map) могут выглядеть следующим образом:

- LongWritable (Входной ключ)
- Text (Входное значение)
- Text (Промежуточный выходной ключ)
- IntWritable (Промежуточное выходное значение)

Четыре параметра функции свёртки (reduce) могут быть такими:

- Text (Промежуточный выходной ключ)
- IntWritable (Промежуточное выходное значение)
- Text (Окончательный выходной ключ)
- IntWritable (Окончательное выходное значение)

7. Для чего предназначен формат файла SequenceFile в Hadoop?

Ответ: Формат файла SequenceFile используется для хранения двоичных пар ключ/значение. Формат SequenceFile позволяет разделять файл даже если данные внутри файла хранятся в сжатом виде, что невозможно при обычном архивировании файлов. Вы можете выбрать как сжатие на уровне записи, при котором отдельные пары ключ/значение будут сжаты. Либо вы можете выбрать сжатие на уровне блока, несколько записей будут сжаты вместе.

8. Что такое платформа Hadoop?

Ответ: Hadoop — это свободно распространяемая платформа, основанная на языке программирования Java и позволяющая проводить вычисления над большими объёмами данных в распределённых кластерах.

Hadoop является частью проекта Apache, который спонсируется фондом Apache Software Foundation. Hadoop позволяет запускать приложения в кластерах, состоящих из тысяч узлов и проводя обработку тысяч терабайт данных. Его распределённая файловая система способствует увеличению скорости передачи данных между узлами и позволяет всему кластеру в целом продолжать работу, даже если один из узлов остановится аварийно. Данный подход снижает риск полного отказа кластера, даже если значительное число узлов выйдут из строя. Hadoop используется многими компаниями, такими как Google, Yahoo и IBM, в основном в приложениях поиска информации и контекстной рекламы. Предпочтительные операционные системы, с которыми работает Hadoop это Linux и Hadoop, но также допустимы как BSD так и OS X.

9. Что такое MapReduce?

Ответ: MapReduce — это модель параллельных вычислений, которая используется для работы с большими объёмами данных в Hadoop кластере, состоящим из сотен или даже тысяч узлов.

Модель MapReduce переносит вычисления к месту расположения данных в отличие от традиционной модели параллелизма, в которой данные переносятся к вычислениям. Вычисления в модели MapReduce состоят из двух фаз: отображения (Map) и свёртки (Reduce). В первой фазе — отображение (Map) — происходит преобразование входного набора данных в выходной, в котором элементы разбиты на кортежи (пары ключ/значение). Во второй фазе — свёртка (Reduce) — происходит обработка результатов предыдущей фазы и преобразование данных также в набор пар ключ/значение, но уже меньшего размера. Как видно из самого названия модели вычислений MapReduce свёртка (Reduce) всегда выполняется после отображения (Map). Язык, на котором реализована модель MapReduce — это Java. Все данные, которые подлежат обработке в модели вычислений MapReduce дожны быть представлены в виде пар ключ/значение.

1 1. Какова структура программы MapReduce?

Ответ: Программа MapReduce состоит из трёх следующих частей:

- Драйвер
- Отображение (Mapper)
- Свёртка (Reducer)

4.3 Промежуточная аттестация по дисциплине проводится в форме зачета

Типовые вопросы зачета (ПК-2)

Типовые вопросы зачета

1 1. Понятие и определение BigData.
2 2. Особенности сбора, хранения, обработки и анализа BigData.
3 3. Требования к распределенным информационным системам.

Типовые задания для зачета (ПК-2)

Типовые задания тестирования

1. Каковы наиболее общие форматы входных данных, определённые в Hadoop? Какой из них используется по умолчанию?

Ответ: Наиболее общие форматы входных данных, определённые в Hadoop это

- TextInputFormat
- KeyValueInputFormat
- SequenceFileInputFormat

По умолчанию используется TextInputFormat.

2. В чём отличие между классами TextInputFormat и KeyValueInputFormat?

Ответ: TextInputFormat читает строки из текстового файла и предоставляет смещение строки относительно начала файла в качестве входного ключа для отображения (Mapper), а саму строку в качестве входного значения для отображения (Mapper).

KeyValueInputFormat читает строки из текстового файла и производит разбор каждой отдельной строки в пару ключ-значение. Все символы от начала строки и до первого символа табуляции передаются отображению (Mapper) в качестве ключа, а остаток передаётся отображению (Mapper) в качестве значения.

3. Что такое трекер заданий (JobTracker)?

Ответ: Трекер заданий (JobTracker) — это сервис платформы Hadoop, который запускает на кластере задания (Job), построенные согласно модели вычисления MapReduce. Трекером заданий (JobTracker) также называется узел кластера, на котором работает указанный сервис.

4. Что такое Hadoop Streaming?

Ответ: Hadoop Streaming — это общий API, позволяющий программе, написанной теоретически на любом языке, реализовать фазы отображения (Map) и свёртки (Reduce) в платформе Hadoop.

5. Как выполнить проверку работоспособности файловой системы HDFS?

Ответ: Для проверки работоспособности файловой системы HDFS используется утилита FSCK. Она очень удобна для проверки целостности файла, имён блоков и расположения блоков.

Пример. Запуск утилиты FSCK. hdfs fsck /dir/hadoop-test -files -blocks -locations

6. Каковы параметры функций отображений (map) и свёртки (reducer)?

Ответ: Сигнатуры методов отображения и свёртки много говорят о типе входных и выходных данных, которыми работает задание (Job). В предположении что вы используете TextInputFormat, параметры функции отображения (Map) могут выглядеть следующим образом:

- LongWritable (Входной ключ)
- Text (Входное значение)
- Text (Промежуточный выходной ключ)
- IntWritable (Промежуточное выходное значение)

Четыре параметра функции свёртки (reduce) могут быть такими:

- Text (Промежуточный выходной ключ)
- IntWritable (Промежуточное выходное значение)
- Text (Окончательный выходной ключ)
- IntWritable (Окончательное выходное значение)

7. Для чего предназначен формат файла SequenceFile в Hadoop?

Ответ: Формат файла SequenceFile используется для хранения двоичных пар ключ/значение. Формат SequenceFile позволяет разделять файл даже если данные внутри файла хранятся в сжатом виде, что невозможно при обычном архивировании файлов. Вы можете выбрать как сжатие на уровне записи, при котором отдельные пары ключ/значение будут сжаты. Либо вы можете выбрать сжатие на уровне блока, несколько записей будут сжаты вместе.

8. Что такое платформа Hadoop?

Ответ: Hadoop — это свободно распространяемая платформа, основанная на языке программирования Java и позволяющая проводить вычисления над большими объёмами данных в распределённых кластерах.

Hadoop является частью проекта Apache, который спонсируется фондом Apache Software Foundation. Hadoop позволяет запускать приложения в кластерах, состоящих из тысяч узлов и проводя обработку тысяч терабайт данных. Его распределённая файловая система способствует увеличению скорости передачи данных между узлами и позволяет всему кластеру в целом продолжать работу, даже если один из узлов остановится аварийно. Данный подход снижает риск полного отказа кластера, даже если значительное число узлов выйдут из строя. Hadoop используется многими компаниями, такими как Google, Yahoo и IBM, в основном в приложениях поиска информации и контекстной рекламы. Предпочтительные операционные системы, с которыми работает Hadoop это Linux и Hadoop, но также допустимы как BSD так и OS X.

9. Что такое MapReduce?

Ответ: MapReduce — это модель параллельных вычислений, которая используется для работы с большими объёмами данных в Hadoop кластере, состоящим из сотен или даже тысяч узлов.

Модель MapReduce переносит вычисления к месту расположения данных в отличие от традиционной модели параллелизма, в которой данные переносятся к вычислениям. Вычисления в модели MapReduce состоят из двух фаз: отображения (Map) и свёртки (Reduce). В первой фазе — отображение (Map) — происходит преобразование входного набора данных в выходной, в котором элементы разбиты на кортежи (пары ключ/значение). Во второй фазе — свёртка (Reduce) — происходит обработка результатов предыдущей фазы и преобразование данных также в набор пар ключ/значение, но уже меньшего размера. Как видно из самого названия модели вычислений MapReduce свёртка (Reduce) всегда выполняется после отображения (Map). Язык, на котором реализована модель MapReduce — это Java. Все данные, которые подлежат обработке в модели вычислений MapReduce дожны быть представлены в виде пар ключ/значение.

1 1. Какова структура программы MapReduce?

Ответ: Программа MapReduce состоит из трёх следующих частей:

- Драйвер
- Отображение (Mapper)
- Свёртка (Reducer)

4.4. Шкала оценивания промежуточной аттестации

Оценка	Компетенции	Дескрипторы (уровни) – основные признаки освоения (показатели достижения результата)
«зачтено» (50 - 100 баллов)	ПК-2	Знает основные подходы и методы сбора, анализа больших данных, умеет разрабатывать алгоритмы анализа больших данных с учетом требований сервисов и ИТ-систем, использовать результаты анализа данных для разработки перспектив развития организации, владеет способами инструментальной обработки информационных данных и применяет их в профессиональной деятельности, владеет способами размещения информации в информационно-коммуникационной сети Интернет.
«не зачтено» (0 - 49 баллов)	ПК-2	Неудовлетворительные знания о подходах и методах сбора, анализа больших данных, не умеет разрабатывать алгоритмы анализа больших данных с учетом требований сервисов и ИТ систем, не владеет в полном объеме способами инструментальной обработки информационных данных и их применения в профессиональной деятельности.

5. Методические указания для обучающихся по освоению дисциплины (модуля)

5.1 Методические указания по организации самостоятельной работы обучающихся:

Приступая к изучению дисциплины, в первую очередь обучающимся необходимо ознакомиться содержанием рабочей программы дисциплины (РПД), которая определяет содержание, объем, а также порядок изучения и преподавания учебной дисциплины, ее раздела, части.

Для самостоятельной работы важное значение имеют разделы «Объем и содержание дисциплины», «Учебно-методическое и информационное обеспечение дисциплины» и «Материально-техническое обеспечение дисциплины, программное обеспечение, профессиональные базы данных и информационные справочные системы».

В разделе «Объем и содержание дисциплины» указываются все разделы и темы изучаемой дисциплины, а также виды занятий и планируемый объем в академических часах.

В разделе «Учебно-методическое и информационное обеспечение дисциплины» указана рекомендуемая основная и дополнительная литература.

В разделе «Материально-техническое обеспечение дисциплины, программное обеспечение, профессиональные базы данных и информационные справочные системы» содержится перечень профессиональных баз данных и информационных справочных систем, необходимых для освоения дисциплины.

5.2 Рекомендации обучающимся по работе с теоретическими материалами по дисциплине

При изучении и проработке теоретического материала необходимо:

- просмотреть еще раз презентацию лекции в системе MOODLe, повторить законспектированный на лекционном занятии материал и дополнить его с учетом рекомендованной дополнительной литературы;

- при самостоятельном изучении теоретической темы сделать конспект, используя рекомендованные в РПД источники, профессиональные базы данных и информационные справочные системы:

- ответить на вопросы для самостоятельной работы, по теме представленные в пункте 3.2 РПД.

- при подготовке к текущему контролю использовать материалы фонда оценочных средств (ФОС).

5.3 Рекомендации по работе с научной и учебной литературой

Работа с основной и дополнительной литературой является главной формой самостоятельной работы и необходима при подготовке к устному опросу на семинарских занятиях, к дебатам, тестированию, экзамену. Она включает проработку лекционного материала и рекомендованных источников и литературы по тематике лекций.

Конспект лекции должен содержать реферативную запись основных вопросов лекции, в том числе с опорой на размещенные в системе MOODLe презентаци, основных источников и литературы по темам, выводы по каждому вопросу. Конспект может быть выполнен в рамках распечатки выдачи презентаций лекций или в отдельной тетради по предмету. Он должен быть аккуратным, хорошо читаемым, не содержать не относящуюся к теме информацию или рисунки.

Конспекты научной литературы при самостоятельной подготовке к занятиям должны содержать ответы на каждый поставленный в теме вопрос, иметь ссылку на источник информации с обязательным указанием автора, названия и года издания используемой научной литературы. Конспект может быть опорным (содержать лишь основные ключевые позиции), но при этом позволяющим дать полный ответ по вопросу, может быть подробным. Объем конспекта определяется самим студентом.

В процессе работы с основной и дополнительной литературой студент может:

- делать записи по ходу чтения в виде простого или развернутого плана (создавать перечень основных вопросов, рассмотренных в источнике);

- составлять тезисы (цитирование наиболее важных мест статьи или монографии, короткое изложение основных мыслей автора);

- готовить аннотации (краткое обобщение основных вопросов работы);

- создавать конспекты (развернутые тезисы).

5.4. Рекомендации по подготовке к отдельным заданиям текущего контроля

Собеседование предполагает организацию беседы преподавателя со студентами по вопросам практического занятия с целью более обстоятельного выявления их знаний по определенному разделу, теме, проблеме и т.п. Все члены группы могут участвовать в обсуждении, добавлять информацию, дискутировать, задавать вопросы и т.д.

Устный опрос может применяться в различных формах: фронтальный, индивидуальный, комбинированный. Основные качества устного ответа подлежащего оценке:

- правильность ответа по содержанию;

- полнота и глубина ответа;

- сознательность ответа;

- логика изложения материала;

- рациональность использованных приемов и способов решения поставленной учебной задачи;

- своевременность и эффективность использования наглядных пособий и технических средств при ответе;

- использование дополнительного материала;

- рациональность использования времени, отведенного на задание.

Устный опрос может сопровождаться презентацией, которая подготавливается по одному из вопросов практического занятия. При выступлении с презентацией необходимо обращать внимание на такие моменты как:

- содержание презентации: актуальность темы, полнота ее раскрытия, смысловое содержание, соответствие заявленной темы содержанию, соответствие методическим требованиям (цели. ссылки на ресурсы. соответствие содержания и литературы), практическая направленность, соответствие содержания заявленной форме, адекватность использования технических средств учебным задачам, последовательность и логичность презентуемого материала;

- оформление презентации: объем (оптимальное количество), дизайн (читаемость, наличие и соответствие графики и анимации, звуковое оформление, структурирование информации, соответствие заявленным требованиям), оригинальность оформления, эстетика, использование возможности программной среды, соответствие стандартам оформления;

- личностные качества: ораторские способности. соблюдение регламента, эмоциональность, умение ответить на вопросы, систематизированные, глубокие и полные знания по всем разделам программы:

- содержание выступления: логичность изложения материала, раскрытие темы, доступность изложения, эффективность применения средств ИКТ, способы и условия достижения результативности и эффективности для выполнения задач своей профессиональной или учебной деятельности, доказательность принимаемых решений, умение аргументировать свои заключения, выводы.

6. Учебно-методическое и информационное обеспечение дисциплины

6.1 Основная литература:

1. Волкова, Т. В., Насейкина, Л. Ф. Разработка систем распределенной обработки данных : учебно-методическое пособие. - Весь срок охраны авторского права; Разработка систем распределенной обработки данных. - Оренбург: Оренбургский государственный университет, ЭБС АСВ, 2012. - 330 с. - Текст : электронный // IPR BOOKS [сайт]. - URL: http://www.iprbookshop.ru/30127.html

2. Бутаков Н. А., Петров М. В., Насонов Д. Обработка больших данных с Apache Spark : учебно-методическое пособие. - Санкт-Петербург: Университет ИТМО, 2019. - 52 с. - Текст : электронный // ЭБС «Университетская библиотека онлайн» [сайт]. - URL: http://biblioclub.ru/index.php?page=book&id=566771

6.2 Дополнительная литература:

1. Малашонок Г.И., Переславцева О.Н., Рыбаков М.А. Параллельное программирование на OpenMPI Java с приложениями в Math Partner : в 3 ч. : учеб. пособие. - Тамбов: [Издат. дом ТГУ им. Г.Р. Державина], 2014

2. Малявко А. А. Параллельное программирование на основе технологий OpenMP, MPI, CUDA : Учебное пособие для вузов. - испр. и доп; 2-е изд.. - Москва: Юрайт, 2020. - 129 с. - Текст : электронный // ЭБС «ЮРАЙТ» [сайт]. - URL: https://urait.ru/bcode/453248

3. Барский А. Б. Параллельное программирование : монография. - 2-е изд., исправ.. - Москва: Национальный Открытый Университет «ИНТУИТ», 2016. - 346 с. - Текст : электронный // ЭБС «Университетская библиотека онлайн» [сайт]. - URL: http://biblioclub.ru/index.php?page=book&id=578026

4. Ч. 2, 2016. - 77 с.

5. Антонов, А. С. Параллельное программирование с использованием технологии MPI. - 2021-01-23; Параллельное программирование с использованием технологии MPI. - Москва: Интернет-Университет Информационных Технологий (ИНТУИТ), 2016. - 83 с. - Текст : электронный // IPR BOOKS [сайт]. - URL: http://www.iprbookshop.ru/73704.html

6. Левин, М. П. Параллельное программирование с использованием OpenMP : учебное пособие. - 2022-07-28; Параллельное программирование с использованием OpenMP. - Москва: Интернет-Университет Информационных Технологий (ИНТУИТ), Ай Пи Ар Медиа, 2020. - 133 с. - Текст : электронный // IPR BOOKS [сайт]. - URL: http://www.iprbookshop.ru/97572.html

7. Арыков С. Б., Городничев М. А., Щукин Г. А. Параллельное программирование над общей памятью: OpenMP : учебное пособие. - Новосибирск: Новосибирский государственный технический университет, 2019. - 95 с. - Текст : электронный // ЭБС «Университетская библиотека онлайн» [сайт]. - URL: http://biblioclub.ru/index.php?page=book&id=576119

6.3 Иные источники:

1. Java Rush - https://javarush.ru/
2. Национальный Открытый Университет «ИНТУИТ» - http://www.intuit.ru/
3. СКА MahtPartner - http://mathpar.cloud.unihub.ru/
4. Apache Hadoop. http://hadoop.apache.org/ - http://hadoop.apache.org/

7. Материально-техническое обеспечение дисциплины, программное обеспечение, профессиональные базы данных и информационные справочные системы

Для проведения занятий по дисциплине необходимо следующее материально-техническое обеспечение: учебные аудитории для проведения занятий лекционного и семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, помещения для самостоятельной работы.

Учебные аудитории и помещения для самостоятельной работы укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Помещения для самостоятельной работы укомплектованы компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду Университета.

Для проведения занятий лекционного типа используются наборы демонстрационного оборудования, обеспечивающие тематические иллюстрации (проектор, ноутбук, экран/ интерактивная доска).

Лицензионное и свободно распространяемое программное обеспечение:

Java 8 Update 151 Oracle Corporation 27.03.2018 99,73 MB 8.0.1510.12

LibreOffice

Microsoft Office Профессиональный плюс 2007

Microsoft Windows 10

Профессиональные базы данных и информационные справочные системы:

1. Цифровой образовательный ресурс IPR SMART. – URL: http://www.iprbookshop.ru
2. Научная электронная библиотека «КиберЛенинка». – URL: https://cyberleninka.ru
3. Научная электронная библиотека eLIBRARY.ru. – URL: https://elibrary.ru
4. Платформа Springer Link. – URL: https://link.springer.com
5. Президентская библиотека имени Б.Н. Ельцина. – URL: https://www.prlib.ru
6. Российская государственная библиотека. – URL: https://www.rsl.ru
7. Российская национальная библиотека. – URL: http://nlr.ru
8. Тамбовская областная универсальная научная библиотека им. А.С. Пушкина. – URL: http://www.tambovlib.ru
9. Университетская библиотека онлайн: электронно-библиотечная система. – URL: https://biblioclub.ru
10. Федеральное хранилище «Единая коллекция цифровых образовательных ресурсов». – URL: http://school-collection.edu.ru
11. Федеральный портал «Российское образование». – URL: https://www.edu.ru
12. ЭБС «Университетская библиотека онлайн» . – URL: http://www.biblioclub.ru
13. Юрайт: электронно-библиотечная система. – URL: https://urait.ru

Электронная информационно-образовательная среда

https://auth.tsutmb.ru/authorize?response_type=code&client_id=moodle&state=xyz

Взаимодействие преподавателя и студента в процессе обучения осуществляется посредством мультимедийных, гипертекстовых, сетевых, телекоммуникационных технологий, используемых в электронной информационно-образовательной среде университета.