

Министерство науки и высшего образования Российской Федерации

федеральное государственное бюджетное образовательное
учреждение высшего образования
«Курганский государственный университет»
(КГУ)

Кафедра «Безопасность информационных и автоматизированных систем»



УТВЕРЖДАЮ:
Первый проректор

/ Т.Р. Змызгова /

«1» августа 2023 г.

Рабочая программа учебной дисциплины

МЕТОДЫ И ИНСТРУМЕНТЫ АНАЛИЗА БОЛЬШИХ ДАННЫХ

образовательной программы высшего образования –
программы специалитета

10.05.03 — Информационная безопасность автоматизированных систем
Специализация № 5: «Безопасность открытых информационных систем»


Формы обучения: очная

Курган 2023

Рабочая программа дисциплины «Методы и инструменты анализа больших данных» составлена в соответствии с учебными планами по программе специалитета «Информационная безопасность автоматизированных систем» (безопасность открытых информационных систем), утвержденным для очной формы обучения « 30 » июня 2023 года.

Рабочая программа дисциплины одобрена на заседании кафедры «Безопасность информационных и автоматизированных систем» 31 августа 2023 года, протокол № 1

Рабочую программу разработал
доцент кафедры БИАС



Д.И. Дик

Согласовано:

Заведующий
кафедрой БИАС



Д.И. Дик

Начальник
Управления
образовательной деятельности



И.В. Григоренко

Специалист
по учебно-методической работе
Учебно-методического отдела



Г.В. Казанкова

1. ОБЪЕМ ДИСЦИПЛИНЫ

Всего: 4 зачетных единиц трудоемкости (144 академических часа)

Очная форма обучения

Вид учебной работы	На всю дисциплину	Семестр
		11
Аудиторные занятия (контактная работа с преподавателем), всего часов в том числе:	64	64
Лекции	32	32
Лабораторные работы	32	32
Самостоятельная работа, всего часов в том числе:	80	80
Подготовка к зачету	18	18
Другие виды самостоятельной работы (подготовка к лабораторным занятиям и рубежному контролю)	62	62
Вид промежуточной аттестации	зачет	зачет
Общая трудоемкость дисциплины и трудоемкость по семестрам, часов	144	144

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Дисциплина «Методы и инструменты анализа больших данных» относится к части Блока 1, формируемой участниками образовательных отношений Блока 1.

Для освоения дисциплины необходимы компетенции, сформированные в результате изучения дисциплин: «Информатика», «Основы программирования», «Технологии и методы программирования», «Безопасность систем баз данных».

Дисциплина является одной из завершающих дисциплин. Знания и практические навыки, полученные при изучении данного курса, используются при подготовке и защите выпускной квалификационной работы.

3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ

Целью дисциплины является обучение студентов основам технологий анализа больших данных.

Задачи дисциплины:

- получение представления о современных подходах к выполнению анализа больших данных;
- умение выполнять оценку применимости технологий работы с большими данными для построения систем анализа;
- использование инструментов работы с большими данными;

Компетенции, формируемые в результате освоения дисциплины:

- способен обрабатывать и анализировать научно-техническую информацию и результаты исследований (ПК-1);
- способен разрабатывать и анализировать проектные решения по обеспечению безопасности автоматизированных систем (ПК-5);
- способен оценивать эффективность систем защиты информации, функционирующих в открытых информационных системах (ПК-8).

В результате изучения дисциплины обучающийся должен:

знать:

- основы архитектур систем хранения больших данных (для ПК-1);
- языки и фреймворки используемые для работы с большими данными (для ПК-5);
- особенности построения систем анализа больших данных (для ПК-8).

уметь:

- подбирать инструменты для решения конкретных задач анализа больших данных (для ПК-5, ПК-8);
- разворачивать системы анализа больших данных, загружать в них данные и выполнять аналитические выборки (для ПК-5).

владеть:

- методами и инструментами анализа больших данных (для ПК-5).

4. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

4.1. Учебно-тематический план. Очная форма обучения

Номер раздела, темы	Наименование раздела, темы	Количество часов контактной работы с преподавателем		
		Лекции	Лабораторные работы	
Рубеж 1	1	Модели данных		
	2	Введение в большие данные	2	-
	3	Технология MapReduce и Apache Hadoop	2	-
	4	Введение в системы NoSQL	4	8
	5	Хранилища ключ-значение	2	-
		Рубежный контроль 1. Тестирование	4	4
Рубеж 2	6	Столбцовые хранилища данных	2	-
	7	Системы Apache Hive и Apache Impala	2	4
	8	Устойчивые распределённые базы данных (RDD) в Apache Spark	4	8
	9	Потоковая обработка и анализ данных в Apache Kafka	4	4
		Рубежный контроль 2. Тестирование	4	4
Итого:			32	32

4.2. Содержание лекционных занятий

Тема 1. Модели данных

Понятие данных. Типы данных: структурированные, не структурированные, полу структурированные. Системы управления реляционными базами данных: свойства ACID.

Тема 2. Введение в большие данные

Понятие «Большие данные». Технологические вызовы «Больших данных». Характеристики больших данных. Источники. Введение в технологии управления большими данными.

Тема 3. Технология MapReduce и Apache Hadoop

Обзор возможностей и назначения сервиса анализа больших данных Hadoop. Распределенная файловая система HDFS. Устойчивость к сбоям и репликация. Принципы технологии MapReduce. Обработка данных с помощью MapReduce. Технические особенности функционирования MapReduce в Hadoop.

Тема 4. Введение в системы NoSQL

Обзор технологий NoSQL (NewSQL). NoSQL и теорема CAP. Виды NoSQL баз данных: хранилища ключ-значение, столбцовые хранилища, хранилища семейства колонок, графовые базы данных, документные базы данных.

Тема 5. Хранилища ключ-значение

Назначение. Архитектура хранилища ключ-значение Redis. Использование Redis.

Тема 6. Столбцовые хранилища данных

Назначение. Архитектура столбцового хранилища ClickHouse. Работа с данными.

Тема 7. Системы Apache Hive и Apache Impala

Назначение Apache Hive и Apache Impala. Архитектура системы Apache Hive. Язык запросов. Архитектура системы Apache Impala.

Тема 8. Устойчивые распределённые базы данных (RDD) в Apache Spark

Введение в Apache Spark. Архитектура распределенного приложения Spark. Основные концепции Spark. RDD и граф преобразований. Основные этапы обработки данных. Загрузка данных из внешнего хранилища. Изменение размещения данных и количества партиций. Вычисления над данными в Spark. Управление памятью в Apache Spark. DataFrame API и Spark SQL.

Тема 9. Поточковая обработка и анализ данных в Apache Kafka

Обмен сообщениями по типу «публикация/подписка». Системы организации очередей. Сообщения и пакеты. Схемы. Темы и разделы. Производители и потребители. Брокеры и кластеры. Несколько кластеров. Сценарии использования.

4.3. Лабораторные работы

Номер темы	Наименование раздела, темы	Наименование тем лабораторных работ	Норматив времени, час.
3	Технология MapReduce и Apache Hadoop	Развертывание системы распределенных вычислений Hadoop	4
		Реализация MapReduce задачи в системе распределенных вычислений Hadoop	4
5	Хранилища ключ-значение	Установка сервера Redis и работа с данными в Redis	4
6	Столбцовые хранилища данных	Установка сервера ClickHouse и работа с данными в ClickHouse	4
7	Системы Apache Hive и Apache Impala	Установка Apache Hive и работа с данными в Apache Hive	4
		Установка Apache Impala и работа с данными в Apache Impala	4
8	Устойчивые распределённые базы данных (RDD) в Apache Spark	Установка Apache Spark и работа с RDD данными	4
9	Потоковая обработка и анализ данных в Apache Kafka	Установка Apache Kafka и работа с данными в Apache Kafka	4
	Итого		32

5. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

При прослушивании лекций рекомендуется в конспекте отмечать все важные моменты, на которых заостряет внимание преподаватель, в частности

те, которые направлены на качественное выполнение соответствующей лабораторной работы.

Залогом качественного выполнения лабораторных работ является самостоятельная подготовка к ним накануне путем повторения материалов лекций. Преподавателем запланировано на лабораторных занятиях коллективное взаимодействие и разбор конкретных ситуаций, а также обсуждение неясных моментов и ситуаций по лекционному курсу.

Для текущего контроля успеваемости преподавателем используется балльно-рейтинговая система контроля и оценки академической активности. Поэтому настоятельно рекомендуется тщательно прорабатывать материал дисциплины при самостоятельной работе, участвовать во всех формах обсуждения и взаимодействия, как на лекциях, так и на лабораторных занятиях в целях лучшего освоения материала и получения высокой оценки по результатам освоения дисциплины.

Выполнение самостоятельной работы подразумевает самостоятельное изучение разделов дисциплины, подготовку к лабораторным занятиям, к рубежным контролям, подготовку к зачету.

Рекомендуемая трудоемкость самостоятельной работы представлена в таблице:

Рекомендуемый режим самостоятельной работы

Наименование вида самостоятельной работы	Рекомендуемая трудоемкость, акад. час.
Самостоятельное изучение тем дисциплины:	
Модели данных	50
Введение в большие данные	2
Технология MapReduce и Apache Hadoop	2
Введение в системы NoSQL	6
Хранилища ключ-значение	4
Столбцовые хранилища данных	7
Системы Apache Hive и Apache Impala	7
Устойчивые распределённые базы данных (RDD) в Apache Spark	8
Потоковая обработка и анализ данных в Apache Kafka	7
Подготовка к лабораторным работам (по 1 ч. на каждое занятие)	8
Подготовка к рубежным контролям (по 2 часа)	4
Подготовка к зачету	4
Всего:	18
	80

6. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ

6.1. Перечень оценочных средств

1. Балльно-рейтинговая система контроля и оценки академической активности студентов в КГУ (для очной формы обучения)
2. Отчеты студентов по лабораторным работам.
3. Банк тестовых заданий к рубежным контролям № 1, № 2
4. Перечень вопросов к зачету.

6.2. Система балльно-рейтинговой оценки работы студентов по дисциплине

№	Наименование	Содержание					
		Распределение баллов					
1	Распределение баллов за семестры по видам учебной работы, сроки сдачи учебной работы (доводятся до сведения студентов на первом учебном занятии)	11 семестр					
		Вид учебной работы:	Посещение лекций	Выполнение и защита лабораторной работы	Рубежный контроль №1	Рубежный контроль №2	Зачет
		Балльная оценка:	1 ₆ x 16 = 16 ₆	4 ₆ x 8 = 32 ₆	11	11	30
2	Критерий пересчета баллов в традиционную оценку по итогам работы в семестре и зачета	60 и менее баллов – неудовлетворительно; не зачтено; 61...73 – удовлетворительно; зачтено; 74... 90 – хорошо; 91...100 – отлично					
3	Критерии допуска к промежуточной аттестации, возможности получения автоматического зачета по дисциплине, возможность получения бонусных баллов	<p>Для допуска к промежуточной аттестации по дисциплине за семестр обучающийся должен набрать по итогам текущего и рубежного контроля не менее 51 баллов. В случае если обучающийся набрал менее 51 балла, то к аттестационным испытаниям он не допускается.</p> <p>Для получения зачета без проведения процедуры промежуточной аттестации обучающемуся необходимо набрать в ходе текущего и рубежных контролей не менее 61 балла. В этом случае итог балльной оценки, получаемой обучающимся, определяется по количеству баллов, набранных им в ходе текущего и рубежного контролей. При этом, на усмотрение преподавателя, балльная оценка обучающегося может быть повышена за счет получения дополнительных баллов за академическую активность.</p> <p>Обучающийся, имеющий право на получение оценки без проведения процедуры промежуточной аттестации, может повысить ее путем сдачи аттестационного испытания. В случае получения обучающимся на аттестационном испытании 0 баллов итог балльной оценки по дисциплине не снижается.</p> <p>За академическую активность в ходе освоения дисциплины, участие в учебной, научно-исследовательской, спортивной, культурно-творческой и общественной деятельности обучающемуся могут быть начислены дополнительные баллы. Максимальное количество дополнительных баллов за академическую активность составляет 30.</p> <p>Основанием для получения дополнительных баллов являются:</p> <ul style="list-style-type: none"> - выполнение дополнительных заданий по дисциплине; дополнительные баллы начисляются преподавателем; - участие в течение семестра в учебной, научно-исследовательской, спортивной, культурно-творческой и общественной деятельности КГУ. 					

4	<p>Формы и виды учебной работы для неуспевающих (восстановившихся на курсе обучения) студентов для получения недостающих баллов в конце семестра</p>	<p>В случае если к промежуточной аттестации (зачету) набрана сумма менее 51 баллов, обучающемуся необходимо набрать недостающее количество баллов за счет выполнения дополнительных заданий, до конца последней (зачетной) недели семестра.</p> <p>Ликвидация академических задолженностей, возникших из-за разности в учебных планах при переводе или восстановлении, проводится путем выполнения дополнительных заданий, форма и объем которых определяется преподавателем.</p>
---	--	---

6.3. Процедура оценивания результатов освоения дисциплины

Рубежные контроли проводятся в форме письменного тестирования.

Зачет – в форме устного ответа на 2 вопроса. Перечень вопросов преподаватель выдает заранее. Время, отводимое студенту на подготовку вопросов, составляет 1 академический час. Каждый вопрос оценивается в 15 баллов.

Перед проведением каждого рубежного контроля преподаватель прорабатывает со студентами основной материал соответствующих разделов дисциплины в форме краткой лекции-дискуссии.

На каждое тестирование при рубежном контроле студенту отводится 2 академических часа. Баллы студенту выставляются в зависимости от числа правильно выбранных ответов. Варианты тестовых заданий для рубежных контролей состоят из 11 вопросов, по 1 баллу каждый.

Результаты текущего контроля успеваемости и зачета заносятся преподавателем в зачетную ведомость, которая сдается в организационный отдел института в день сдачи зачета, а также выставляются в зачетную книжку студента.

6.4. Примеры оценочных средств для рубежных контролей и зачета

1-ый рубежный контроль

1. Что значит I в аббревиатуре ACID?
 - а) Incomplete
 - б) Indefinite
 - в) Isolated
2. Какие из указанных компонентов входят в состав YARN?
 - а) Node Manager
 - б) Resource Manager
 - в) Journal Node
 - г) Resource Tracker
 - д) Scheduler
 - е) Application Manager
3. Замена 12 ядерного сервера с 64 Гб памяти на 32 ядерный сервер с 128 Гб памяти является примером
 - а) вертикального масштабирования
 - б) горизонтального масштабирования
 - в) не верно ни то, ни другое

2-ой рубежный контроль

1. Как расшифровывается RDD?
 - a) Rigid Data Delivery
 - б) Reduced Data Distributer
 - в) Resilient Distributed Dataset
 - г) Redundant District Dataset
 - д) Regularized Derived Dataset
2. Как расшифровывается DStream?
 - a) Data Stream
 - б) Discretized Stream
 - в) Decomposed Stream
 - г) Distributed Stream
 - д) Deterministic Stream
3. Какой из указанных компонентов входят в состав Hive?
 - a) Data Node
 - б) Driver
 - в) Job Tracker

Примерный перечень вопросов к зачету

1. Типы данных: структурированные, не структурированные, полуструктурированные.
2. Характеристики ACID баз данных.
3. Характеристики больших данных и их источники.
4. Распределенная файловая система HDFS. Устойчивость к сбоям и репликация.
5. Принципы технологии MapReduce. Обработка данных с помощью MapReduce.
6. Технические особенности функционирования MapReduce в Hadoop.
7. Обзор технологий NoSQL (NewSQL). NoSQL и теорема CAP.
8. Виды NoSQL баз данных: хранилища ключ-значение, столбцовые хранилища, хранилища семейства колонок, графовые базы данных, документные базы данных.
9. Архитектура хранилища ключ-значение Redis.
10. Архитектура столбцового хранилища ClickHouse.
11. Назначение Apache Hive и Apache Impala.
12. Назначение и архитектура системы Apache Hive.
13. Язык запросов Apache Hive.
14. Назначение и архитектура системы Apache Impala.
15. Архитектура распределенного приложения Spark. Основные концепции Spark.
16. RDD и граф преобразований. Основные этапы обработки данных.
17. Обмен сообщениями по типу «публикация/подписка». Системы организации очередей.
18. Назначение и архитектура системы Apache Kafka.

6.5. Фонд оценочных средств

Полный банк заданий для текущего, рубежных контролей и промежуточной аттестации по дисциплине, показатели, критерии, шкалы оценивания компетенций, методические материалы, определяющие процедуры оценивания образовательных результатов, приведены в учебно-методическом комплексе дисциплины.

7. ОСНОВНАЯ И ДОПОЛНИТЕЛЬНАЯ УЧЕБНАЯ ЛИТЕРАТУРА

7.1. Основная учебная литература

1 Ёсу, М. Т. Принципы организации распределенных баз данных [Электронный ресурс] / М. Т. Ёсу, П. Вальдурис; пер. с англ. А. А. Слинкина. – Электрон. тестовые дан. – М.: ДМК Пресс, 2021. – 672 с. – Доступ из ЭБС «Консультант студента».

7.2. Дополнительная учебная литература

1 Редмонд, Э. Семь баз данных за семь недель. Введение в современные базы данных и идеологию NoSQL [Электронный ресурс] / Эрик Редмонд, Джим. Р. Уилсон; пер. с англ. А. А. Слинкин. – Электрон. тестовые дан. – М.: ДМК Пресс, 2013. – 384 с. – Доступ из ЭБС «Консультант студента».

2 Пселтис, Эндрю Дж. Поточная обработка данных. Конвейер реального времени [Электронный ресурс] / Эндрю Дж. Пселтис; пер. с англ. А. А. Слинкин. – Электрон. тестовые дан. – М.: ДМК Пресс, 2018. – 218 с. – Доступ из ЭБС «Консультант студента».

3 Лэм, Ч. Надоор в действии [Электронный ресурс] / Чак Лэм. – Электрон. тестовые дан. – М.: ДМК Пресс, 2012. – 424 с. – Доступ из ЭБС «Консультант студента».

4 Карау, Х. Изучаем Spark: молниеносный анализ данных [Электронный ресурс] / Х. Карау, Э. Конвински, П. Венделл, М. Захария. – Электрон. тестовые дан. – М.: ДМК Пресс, 2015. – 304 с. – Доступ из ЭБС «Консультант студента».

8. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ ОБУЧАЮЩИХСЯ

1. Дик Д.И. Лабораторный практикум по дисциплине «Методы и инструменты анализа больших данных» для студентов программы специалитета 10.05.03 «Информационная безопасность автоматизированных систем» [Электронный ресурс] / Д.И. Дик. – Электрон. текстовые дан. – Курган: КГУ, 2021. – 100 с.

9. ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ, ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ И ИНФОРМАЦИОННЫЕ СПРАВОЧНЫЕ СИСТЕМЫ

1. ЭБС «Лань».
2. ЭБС «Консультант студента».
3. ЭБС «Znanium.com».

4. «Гарант» - справочно-правовая система.

10. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Материально-техническое обеспечение по реализации дисциплины осуществляется в соответствии с требованиями ФГОС ВО по данной образовательной программе.

11. Для студентов, обучающихся с использованием дистанционных образовательных технологий

При использовании электронного обучения и дистанционных образовательных технологий (далее ЭО и ДОТ) занятия полностью или частично проводятся в режиме онлайн. Объем дисциплины и распределение нагрузки по видам работ соответствует п. 4.1. Распределение баллов соответствует п. 6.2 либо может быть изменено в соответствии с решением кафедры, в случае перехода на ЭО и ДОТ в процессе обучения. Решение кафедры об используемых технологиях и системе оценивания достижений, обучающихся принимается с учетом мнения ведущего преподавателя и доводится до сведения обучающихся.

Аннотация к рабочей программе дисциплины
«Методы и инструменты анализа больших данных»

образовательной программы высшего образования –
программы специалитета

10.05.03 – Информационная безопасность автоматизированных систем
Специализация № 5 «Безопасность открытых информационных систем»

Трудоемкость дисциплины: 4 з.е. (144 академических часа)

Семестр: 11

Форма промежуточной аттестации: зачет.

Содержание дисциплины

Модели данных. Введение в большие данные. Технология MapReduce и Apache Hadoop. Введение в системы NoSQL. Хранилища ключ-значение. Столбцовые хранилища данных. Системы Apache Hive и Apache Impala. Устойчивые распределённые базы данных (RDD) в Apache Spark. Поточная обработка и анализ данных в Apache Kafka.