

**МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ**  
**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ**  
**УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**  
**«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»**  
**(РУТ (МИИТ))**



Рабочая программа дисциплины (модуля),  
как компонент образовательной программы  
высшего образования - программы магистратуры  
по направлению подготовки  
09.04.01 Информатика и вычислительная техника,  
утвержденной первым проректором РУТ (МИИТ)  
Тимониным В.С.

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)**

**Сбор, хранение и обработка больших данных**

Направление подготовки: 09.04.01 Информатика и вычислительная техника

Направленность (профиль): Искусственный интеллект и предиктивная аналитика в транспортных системах

Форма обучения: Очная

Рабочая программа дисциплины (модуля) в виде  
электронного документа выгружена из единой  
корпоративной информационной системы управления  
университетом и соответствует оригиналу

Простая электронная подпись, выданная РУТ (МИИТ)  
ID подписи: 5665  
Подписал: заведующий кафедрой Нутович Вероника  
Евгеньевна  
Дата: 01.09.2024

## 1. Общие сведения о дисциплине (модуле).

Целью освоения данной дисциплины является получение базовых, теоретических знаний и навыков в области сбора, очистки, подготовки, разметки данных для дальнейшего обучения аналитических моделей и моделей искусственного интеллекта.

Задачей освоения дисциплины является формирование базовых представлений и знаний о работе распределенной файловой системы, основных моделей обработки и подготовки данных, основы пакетной обработки данных и обработки данных в реальном времени.

## 2. Планируемые результаты обучения по дисциплине (модулю).

Перечень формируемых результатов освоения образовательной программы (компетенций) в результате обучения по дисциплине (модулю):

**ОПК-6** - Способен разрабатывать компоненты программно-аппаратных комплексов обработки информации и автоматизированного проектирования;

**ПК-2** - Способен осуществить сбор, очистку, подготовку и разметку данных используя методологию ETL для дальнейшего обучения моделей искусственного интеллекта.

Обучение по дисциплине (модулю) предполагает, что по его результатам обучающийся будет:

### **Уметь:**

- разрабатывать задачи обработки больших данных в парадигме MapReduce;
- разрабатывать запросы в Hive для обработки больших данных;
- разрабатывать запросы в Spark Dataframe API для обработки больших данных;
- разворачивать инструменты сбора, хранения и обработки больших данных.

### **Знать:**

- особенности работы распределенной файловой системой;
- основные модели обработки данных;
- основы пакетной обработки данных;
- диалект HiveQL;
- основы итеративной обработки больших данных на Apache Spark;
- основы обработки данных в реальном времени.

### **Владеть:**

- навыками разработки запросов в режиме пакетной обработки данных и режиме обработки данных в реальном времени;
- навыками работы с распределенными файловыми системами;
- навыками работы с Hadoop и Hive для пакетной обработки данных;
- навыками планирования задач в MapReduce;
- навыками работы с Spark Streaming и Kafka для обработки данных в реальном времени.

### 3. Объем дисциплины (модуля).

#### 3.1. Общая трудоемкость дисциплины (модуля).

Общая трудоемкость дисциплины (модуля) составляет 4 з.е. (144 академических часа(ов)).

3.2. Объем дисциплины (модуля) в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении учебных занятий:

Тип учебных занятий	Количество часов	
	Всего	Семестр №1
Контактная работа при проведении учебных занятий (всего):	32	32
В том числе:		
Занятия лекционного типа	16	16
Занятия семинарского типа	16	16

3.3. Объем дисциплины (модуля) в форме самостоятельной работы обучающихся, а также в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении промежуточной аттестации составляет 112 академических часа (ов).

3.4. При обучении по индивидуальному учебному плану, в том числе при ускоренном обучении, объем дисциплины (модуля) может быть реализован полностью в форме самостоятельной работы обучающихся, а также в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении промежуточной аттестации.

### 4. Содержание дисциплины (модуля).

#### 4.1. Занятия лекционного типа.

№ п/п	Тематика лекционных занятий / краткое содержание
1	<p>Введение в «Большие данные».</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- введение в понятие «данные», «информация» и «знания»;</li> <li>- введение в понятие «большие данные»;</li> <li>- признаки больших данных;</li> <li>- особенности инструментов работы с большими данными;</li> <li>- виды обработки, пакетная и в реальном времени;</li> <li>- примеры кейсов и задач обработки больших данных.</li> </ul>
2	<p>Введение в Hadoop. Хранение данных в распределенных файловых системах.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- введение в Hadoop, основные вехи развития;</li> <li>- распределенные файловые системы, основы GFS, HDFS;</li> <li>- форматы хранения данных в HDFS;</li> <li>- работа в HDFS с помощью Java API, терминала и Python.</li> </ul>
3	<p>Введение в Hadoop. Пакетная обработка данных в Hadoop.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- вычислительное ядро Hadoop;</li> <li>- концепция вычислений MapReduce;</li> <li>- архитектура MapReduce;</li> <li>- архитектура YARN.</li> </ul>
4	<p>Базы данных и Hadoop. Озеро данных.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- SQL и NoSQL базы данных;</li> <li>- SQL и NoSQL инструменты в Hadoop;</li> <li>- введение в Hive, архитектура и модель данных в Hive;</li> <li>- диалект HiveQL, обработка данных в Hive;</li> <li>- введение в понятие «озеро данных»;</li> <li>- принципы подходы к построению корпоративного озера данных;</li> <li>- конвейеры обработки данных;</li> <li>- принципы управления большими данными.</li> </ul>
5	<p>Введение в Apache Spark.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- введение в итеративную обработку больших данных на Apache Spark;</li> <li>- особенности обработки больших данных на Apache Spark, отличие от MapReduce;</li> <li>- Spark RDD API;</li> <li>- Spark Dataframe API;</li> <li>- SQL-запросы на Spark;</li> <li>- GraphX и GraphFrames.</li> </ul>
6	<p>Введение в Apache Spark. Обработка данных в реальном времени.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- принципы обработки данных в реальном времени;</li> <li>- обзор возможностей в Spark Streaming API;</li> <li>- обработка данных в реальном времени с помощью Spark Streaming API.</li> </ul>
7	<p>Введение в Kafka. Распределенный брокер сообщений.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- область применения и примеры использования;</li> </ul>

№ п/п	Тематика лекционных занятий / краткое содержание
	<ul style="list-style-type: none"> <li>- компоненты и архитектура Apache Kafka;</li> <li>- брокеры, поставщики и потребители данных, работа с сообщениями;</li> <li>- базовые операции Apache Kafka;</li> <li>- сценарии интеграции с Apache Kafka;</li> <li>- общие понятия Kafka Stream, работа с потоками;</li> <li>- обработка данных в в реальном времени из Kafka.</li> </ul>
8	<p>Развертывание инфраструктуры для сбора, хранения и обработки больших данных.</p> <p>Рассматриваемые вопросы:</p> <ul style="list-style-type: none"> <li>- развертывание Hadoop и его инструментов;</li> <li>- настройка Hive и HBase;</li> <li>- настройка Apache Spark;</li> <li>- развертывание Kafka.</li> </ul>

## 4.2. Занятия семинарского типа.

### Лабораторные работы

№ п/п	Наименование лабораторных работ / краткое содержание
1	<p>Сбор и хранение данных в распределенных файловых системах.</p> <p>В результате выполнения лабораторной работы студент получает навык работы с HDFS для сбора и хранения данных.</p>
2	<p>Введение в MapReduce. Обработка данных с помощью MapReduce.</p> <p>В результате выполнения лабораторной работы студент знакомится с концепцией MapReduce для обработки больших данных и получает навык обработки больших данных с использованием MapReduce.</p>
3	<p>Введение в Hive.</p> <p>В результате выполнения лабораторной работы знакомится с диалектом HiveQL.</p>
4	<p>Обработка данных с помощью Hive.</p> <p>В результате выполнения лабораторной работы студент получает навык обработки больших данных с использованием Hive.</p>
5	<p>Введение в Apache Spark.</p> <p>В результате выполнения лабораторной работы студент знакомится с итеративной обработкой больших данных на Apache Spark.</p>
6	<p>Обработка данных с помощью Apache Spark.</p> <p>В результате выполнения лабораторной работы студент получает навык обработки больших данных с использованием Apache Spark.</p>
7	<p>Введение в Apache Kafka.</p> <p>В результате выполнения лабораторной работы студент знакомится с распределенным брокером сообщений Apache Kafka.</p>
8	<p>Обработка данных с помощью Apache Kafka и Spark Streaming API.</p> <p>В результате выполнения лабораторной работы студент получает навык обработки больших данных в реальном времени с помощью Apache Kafka и Spark Streaming API.</p>

## 4.3. Самостоятельная работа обучающихся.

№ п/п	Вид самостоятельной работы
1	Изучение рекомендованной литературы.

№ п/п	Вид самостоятельной работы
2	Подготовка к практическим работам.
3	Выполнение курсового проекта.
4	Подготовка к промежуточной аттестации.
5	Подготовка к текущему контролю.

#### 4.4. Примерный перечень тем курсовых проектов

- 1.Пакетная обработка текстовых данных с помощью MapReduce.
- 2.Пакетная обработка логов с помощью MapReduce.
- 3.Пакетная обработка данных об аренде автомобилей с помощью MapReduce.
- 4.Пакетная обработка данных результатов футбольных матчей с помощью MapReduce.
- 5.Пакетная обработка данных о погоде с помощью Hive.
- 6.Пакетная обработка данных о пассажирских перевозках с помощью Hive.
- 7.Обработка в реальном времени данных об аренде самокатов с помощью Spark Streaming.
- 8.Обработка в реальном времени данных об автомобильном трафике с помощью Spark Streaming.
- 9.Обработка в реальном времени данных сообщений пользователей с помощью Spark Streaming.
- 10.Обработка в реальном времени данных в виде логов с помощью Spark Streaming.

#### 5. Перечень изданий, которые рекомендуется использовать при освоении дисциплины (модуля).

№ п/п	Библиографическое описание	Место доступа
1	Чак, Л. Nadoop в действии / Л. Чак. — Москва : ДМК Пресс, 2012. — 424 с. — ISBN 978-5-94074-785-7.	<a href="https://e.lanbook.com/book/39997">https://e.lanbook.com/book/39997</a> (дата обращения: 05.11.2022г.)
2	Макшанов, А. В. Большие данные. Big Data / А. В. Макшанов, А. Е. Журавлев, Л. Н. Тындыкарь. — 2-е изд., стер. — Санкт-Петербург : Лань, 2022. — 188 с. — ISBN 978-5-8114-9690-7.	<a href="https://e.lanbook.com/book/198599">https://e.lanbook.com/book/198599</a> (дата обращения: 05.11.2022г.)

3	Лебедев, А. С. Методы Big Data : учебно-методическое пособие / А. С. Лебедев, Ш. Г. Магомедов. — Москва : РТУ МИРЭА, 2021. — 91 с.	<a href="https://e.lanbook.com/book/182452">https://e.lanbook.com/book/182452</a> (дата обращения: 05.11.2022г.)
4	Перрен, Ж. -. Spark в действии / Ж. -. Перрен ; перевод с английского А. В. Снастина. — Москва : ДМК Пресс, 2021. — 636 с. — ISBN 978-5-97060-879-1.	<a href="https://e.lanbook.com/book/241001">https://e.lanbook.com/book/241001</a> (дата обращения: 05.11.2022г.)
5	Пселтис, Э. Д. Потокковая обработка данных. Конвейер реального времени / Э. Д. Пселтис ; перевод с английского А. А. Слинкин. — Москва : ДМК Пресс, 2018. — 218 с. — ISBN 978-5-97060-606-3.	<a href="https://e.lanbook.com/book/105840">https://e.lanbook.com/book/105840</a> (дата обращения: 05.11.2022г.)

6. Перечень современных профессиональных баз данных и информационных справочных систем, которые могут использоваться при освоении дисциплины (модуля).

Электронно-библиотечная система издательства «Лань»  
(<http://e.lanbook.com/>)

Электронно-библиотечная система ibooks.ru (<http://ibooks.ru/>)

Открытые лекции (<https://sphere.vk.company/materials/video/#19>)

7. Перечень лицензионного и свободно распространяемого программного обеспечения, в том числе отечественного производства, необходимого для освоения дисциплины (модуля).

Браузер Microsoft Internet Explorer или его аналоги

Пакет офисных программ Microsoft Office или его аналоги

Python 3.6 и выше

Hadoop

Hive

Apache Spark

Apache Kafka

8. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю).

Учебные аудитории для проведения учебных занятий, оснащенные компьютерной техникой и наборами демонстрационного оборудования.

Для практических занятий — наличие персональных компьютеров вычислительного класса.

9. Форма промежуточной аттестации:

Курсовой проект в 1 семестре.

Экзамен в 1 семестре.

10. Оценочные материалы.

Оценочные материалы, применяемые при проведении промежуточной аттестации, разрабатываются в соответствии с локальным нормативным актом РУТ (МИИТ).



Авторы:

заведующий кафедрой, доцент, к.н.  
кафедры «Цифровые технологии  
управления транспортными  
процессами»

В.Е. Нутович

старший преподаватель кафедры  
«Цифровые технологии управления  
транспортными процессами»

Е.А. Заманов

Согласовано:

Заведующий кафедрой ЦТУТП

В.Е. Нутович

Председатель учебно-методической  
комиссии

Н.А. Андриянова