

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



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

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)**

**Распределенные хранилища данных**

Направление подготовки: 09.04.01 Информатика и вычислительная техника

Направленность (профиль): Технологии проектирования программного обеспечения

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 ФГАОУ ВО РУТ  
(МИИТ)

ID подписи: 5665

Подписал: заведующий кафедрой Нутович Вероника  
Евгеньевна

Дата: 01.09.2026

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

Целью освоения данной дисциплины является получение базовых, теоретических знаний и навыков в области построения распределенных хранилищ данных.

Задачи освоения дисциплины:

- формирование у обучающихся базовых представлений о работе распределенной файловой системы, основы пакетной обработки данных и обработки данных в реальном времени;
- формирование у обучающихся навыков разработки запросов в режиме пакетной обработки данных и обработки данных в реальном времени, работы с Hadoop, Spark и Kafka.

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

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

**ПК-2** - Способен проектировать и разрабатывать распределенные высокопроизводительные программные продукты с применением методов оптимизации программного обеспечения для корпоративного рынка.

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 10 з.е. (360 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 |    |
|                                                           |                  | №2      | №3 |
| Контактная работа при проведении учебных занятий (всего): | 96               | 48      | 48 |
| В том числе:                                              |                  |         |    |
| Занятия лекционного типа                                  | 32               | 16      | 16 |
| Занятия семинарского типа                                 | 64               | 32      | 32 |

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 264 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.

| №<br>п/п | Тематика лекционных занятий / краткое содержание                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1        | <b>Введение в «Распределенные данные».</b><br>Рассматриваемые вопросы: <ul style="list-style-type: none"><li>- введение в понятие «данные», «информация» и «знания»;</li><li>- введение в понятие «большие данные» и «распределенные данные»;</li><li>- признаки больших данных;</li><li>- особенности инструментов работы с большими и распределенными данными;</li><li>- виды обработки, пакетная и в реальном времени;</li><li>- примеры кейсов и задач обработки больших данных.</li></ul> |
| 2        | <b>Введение в Hadoop. Хранение данных в распределенных файловых системах.</b><br>Рассматриваемые вопросы: <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        | <b>Введение в Hadoop. Пакетная обработка данных в Hadoop.</b><br>Рассматриваемые вопросы: <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        | <b>Базы данных и Hadoop.</b><br>Рассматриваемые вопросы: <ul style="list-style-type: none"><li>- SQL и NoSQL базы данных;</li><li>- SQL и NoSQL инструменты в Hadoop.</li></ul>                                                                                                                                                                                                                                                                                                                |
| 5        | <b>Озеро данных.</b><br>Рассматриваемые вопросы: <ul style="list-style-type: none"><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>                                                                                                                                                                                               |
| 6        | <b>Введение в Apache Spark.</b><br>Рассматриваемые вопросы: <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>                                                                                                   |
| 7        | <b>Введение в Apache Spark. Обработка данных в реальном времени.</b><br>Рассматриваемые вопросы: <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>                                                                                                                                                              |
| 8        | <b>Введение в Kafka. Распределенный брокер сообщений.</b><br>Рассматриваемые вопросы: <ul style="list-style-type: none"><li>- область применения и примеры использования;</li><li>- компоненты и архитектура Apache Kafka;</li></ul>                                                                                                                                                                                                                                                           |

| №<br>п/п | Тематика лекционных занятий / краткое содержание                                                                                                                                                                                                                                                                                  |
|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|          | <ul style="list-style-type: none"> <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> |
| 9        | <b>Развертывание инфраструктуры для сбора, хранения и обработки больших данных.</b><br>Рассматриваемые вопросы: <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>                |
| 10       | <b>Введение в Hive.</b><br>Рассматриваемые вопросы: <ul style="list-style-type: none"> <li>- введение в Hive, архитектура и модель данных в Hive;</li> <li>- диалект HiveQL, обработка данных в Hive.</li> </ul>                                                                                                                  |

## 4.2. Занятия семинарского типа.

### Практические занятия

| №<br>п/п | Тематика практических занятий/краткое содержание                                                                                                                                                                                      |
|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1        | <b>Сбор и хранение данных в распределенных файловых системах.</b><br>В результате выполнения практической работы студент получает навык работы с HDFS для сбора и хранения данных.                                                    |
| 2        | <b>Введение в MapReduce.</b><br>В результате выполнения практической работы студент знакомится с концепцией MapReduce для обработки больших данных.                                                                                   |
| 3        | <b>Обработка данных с помощью MapReduce.</b><br>В результате выполнения практической работы студент получает навык обработки больших данных с использованием MapReduce.                                                               |
| 4        | <b>Введение в Hive.</b><br>В результате выполнения практической работы знакомится с диалектом HiveQL.                                                                                                                                 |
| 5        | <b>Обработка данных с помощью Hive.</b><br>В результате выполнения практической работы студент получает навык обработки больших данных с использованием Hive.                                                                         |
| 6        | <b>Введение в Apache Spark.</b><br>В результате выполнения практической работы студент знакомится с итеративной обработкой больших данных на Apache Spark.                                                                            |
| 7        | <b>Обработка данных с помощью Apache Spark.</b><br>В результате выполнения практической работы студент получает навык обработки больших данных с использованием Apache Spark.                                                         |
| 8        | <b>Введение в Apache Kafka.</b><br>В результате выполнения практической работы студент знакомится с распределенным брокером сообщений Apache Kafka.                                                                                   |
| 9        | <b>Обработка данных с помощью Apache Kafka и Spark Streaming API.</b><br>В результате выполнения практической работы студент получает навык обработки больших данных в реальном времени с помощью Apache Kafka и Spark Streaming API. |

#### 4.3. Самостоятельная работа обучающихся.

| №<br>п/п | Вид самостоятельной работы             |
|----------|----------------------------------------|
| 1        | Изучение рекомендованной литературы.   |
| 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).

| №<br>п/п | Библиографическое описание                                                                                                                      | Место доступа                                                                                                     |
|----------|-------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|
| 1        | Чак, Л. Надоор в действии / Л. Чак. — Москва : ДМК Пресс, 2012. — 424 с. — ISBN 978-5-94074-785-7. — Текст : электронный                        | <a href="https://e.lanbook.com/book/39997">https://e.lanbook.com/book/39997</a><br>(дата обращения: 17.04.2025)   |
| 2        | Макшанов, А. В. Большие данные. Big Data / А. В. Макшанов, А. Е. Журавлев, Л. Н. Тындыкарь. — 4-е изд., стер. — Санкт-Петербург : Лань, 2024. — | <a href="https://e.lanbook.com/book/362318">https://e.lanbook.com/book/362318</a><br>(дата обращения: 17.04.2025) |

|   |                                                                                                                                                                                                                    |                                                                                                                   |
|---|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|
|   | 188 с. — ISBN 978-5-507-47346-5. — Текст : электронный                                                                                                                                                             |                                                                                                                   |
| 3 | Лебедев, А. С. Методы Big Data : учебно-методическое пособие / А. С. Лебедев, Ш. Г. Магомедов. — Москва : РТУ МИРЭА, 2021. — 91 с. — Текст : электронный Учебно-методическое издание                               | <a href="https://e.lanbook.com/book/182452">https://e.lanbook.com/book/182452</a><br>(дата обращения: 17.04.2025) |
| 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. — Текст : электронный                                         | <a href="https://e.lanbook.com/book/241001">https://e.lanbook.com/book/241001</a><br>(дата обращения: 17.04.2025) |
| 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. — Текст : электронный | <a href="https://e.lanbook.com/book/105840">https://e.lanbook.com/book/105840</a><br>(дата обращения: 17.04.2025) |

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

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

Научно-техническая библиотека РУТ (МИИТ) (<http://library.miit.ru>).

Образовательная платформа «юрайт» (<https://urait.ru/>).

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

Браузер с доступом в интернет

Пакет офисных приложений

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:

Зачет во 2 семестре.

Курсовой проект в 3 семестре.

Экзамен в 3 семестре.

10. Оценочные материалы.

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

Авторы:

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