**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 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

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):

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

**ПК-3** - Способен проектировать, разрабатывать, обучать, оценивать и разворачивать модели искусственного интеллекта в соответствии с DevOps, DataOps и MLOps методологиями.

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;
- развертывать 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 5 з.е. (180 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): | 48 | 48 |
| В том числе: | | |
| Занятия лекционного типа | 16 | 16 |
| Занятия семинарского типа | 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 132 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 | Введение в «Большие данные».<br>Рассматриваемые вопросы:<br>- введение в понятие «данные», «информация» и «знания»;<br>- введение в понятие «большие данные»;<br>- признаки больших данных;<br>- особенности инструментов работы с большими данными;<br>- виды обработки, пакетная и в реальном времени;<br>- примеры кейсов и задач обработки больших данных. |
| 2 | Введение в Hadoop. Хранение данных в распределенных файловых системах.<br>Рассматриваемые вопросы:<br>- введение в Hadoop, основные вехи развития;<br>- распределенные файловые системы, основы GFS, HDFS;<br>- форматы хранения данных в HDFS;<br>- работа в HDFS с помощью Java API, терминала и Python. |
| 3 | Введение в Hadoop. Пакетная обработка данных в Hadoop.<br>Рассматриваемые вопросы:<br>- вычислительное ядро Hadoop;<br>- концепция вычислений MapReduce;<br>- архитектура MapReduce;<br>- архитектура YARN. |
| 4 | Базы данных и Hadoop и Озеро данных.<br>Рассматриваемые вопросы:<br>- SQL и NoSQL базы данных;<br>- SQL и NoSQL инструменты в Hadoop;<br>- введение в Hive, архитектура и модель данных в Hive;<br>- диалект HiveQL, обработка данных в Hive;<br>- введение в понятие «озеро данных»;<br>- приницпыи подходы к построению корпоративного озера данных;<br>- конвейеры обработки данных;<br>- принципы управления большими данными. |
| 5 | Введение в Apache Spark.<br>Рассматриваемые вопросы:<br>- введение в итеративную обработку больших данных на Apache Spark;<br>- особенности обработки больших данных на Apache Spark, отличие от MapReduce;<br>- Spark RDD API;<br>- Spark Dataframe API;<br>- SQL-запросы на Spark;<br>- GraphX и GraphFrames. |
| 6 | Введение в Apache Spark. Обработка данных в реальном времени.<br>Рассматриваемые вопросы:<br>- принципы обработки данных в реальном времени;<br>- обзор возможностей в Spark Streaming API;<br>- обработка данных в реальном времени с помощью Spark Streaming API. |
| 7 | Введение в Kafka. Распределенный брокер сообщений.<br>Рассматриваемые вопросы:<br>- область применения и примеры использования;<br>- компоненты и архитектура Apache Kafka; |

| № п/п | Тематика лекционных занятий / краткое содержание |
|---|---|
| | - брокеры, поставщики и потребители данных, работа с сообщениями;<br>- базовые операции Apache Kafka;<br>- сценарии интеграции с Apache Kafka;<br>- общие понятия Kafka Stream, работа с потоками;<br>- обработка данных в в реальном времени из Kafka. |
| 8 | Развертывание инфраструктуры для сбора, хранения и обработки больших данных.<br>Рассматриваемые вопросы:<br>- развертывание Hadoop и его инструментов;<br>- настройка Hive и HBase;<br>- настройка Apache Spark;<br>- развертывание Kafka. |

## 4.2. Занятия семинарского типа.

### Лабораторные работы

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

## 4.3. Самостоятельная работа обучающихся.

| № п/п | Вид самостоятельной работы |
|---|---|
| 1 | Изучение рекомендованной литературы. |
| 2 | Подготовка к лабораторным 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. Hadoop в действии / Л. Чак. — Москва : ДМК Пресс, 2012. — 424 с. — ISBN 978-5-94074-785-7. — Текст : электронный | https://e.lanbook.com/book/39997 (дата обращения: 11.04.2025) |
| 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. — 188 с. — ISBN 978-5-507-47346-5. — Текст : электронный | https://e.lanbook.com/book/362318 (дата обращения: 11.04.2025) |
| 3 | Лебедев, А. С. Методы Big Data : учебно-методическое пособие / А. С. Лебедев, Ш. Г. Магомедов. — Москва : РТУ МИРЭА, 2021. — 91 с. — Текст : электронный | https://e.lanbook.com/book/182452 (дата обращения: 11.04.2025) |

| | | |
|---|---|---|
| 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 | https://e.lanbook.com/book/241001 (дата обращения: 11.04.2025) |
| 5 | Пселтис, Э. Д. Потоковая 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 | https://e.lanbook.com/book/105840 (дата обращения: 11.04.2025) |

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

- Научно-техническая библиотека РУТ (МИИТ) (http://library.miit.ru);
- Информационный портал Научная электронная библиотека eLIBRARY.RU (www.elibrary.ru );
- Образовательная платформа «Юрайт» (https://urait.ru/);
- Электронно-библиотечная система издательства «Лань» (http://e.lanbook.com/);

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 лабораторных работ – 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