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)

Распределенные хранилища данных

Направление подготовки: 09.04.01 Информатика и вычислительная техника

Направленность (профиль): Технологии проектирования программного обеспечения

Форма обучения: Очная

1. Общие сведения о дисциплине (модуле).

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

В рамках дисциплины у обучающихся формируются базовые представления и знания о работе распределенной файловой системы, основы пакетной обработки данных и обработки данных в реальном времени.

На практических занятиях у обучающихся формируются навыки 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):

ОПК-5 - Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем;

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

ОПК-7 - Способен адаптировать зарубежные комплексы обработки информации и автоматизированного проектирования к нуждам отечественных предприятий;

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;
- 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 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.

№ п/п	Тематика лекционных занятий / краткое содержание
1	Введение в «Распределенные данные». Рассматриваемые вопросы: - введение в понятие «данные», «информация» и «знания»; - введение в понятие «большие данные» и «распределенные данные»; - признаки больших данных; - особенности инструментов работы с большими и распределенными данными; - виды обработки, пакетная и в реальном времени; - примеры кейсов и задач обработки больших данных.
2	Введение в Hadoop. Хранение данных в распределенных файловых системах. Рассматриваемые вопросы: - введение в Hadoop, основные вехи развития; - распределенные файловые системы, основы GFS, HDFS; - форматы хранения данных в HDFS; - работа в HDFS с помощью Java API, терминала и Python.
3	Введение в Hadoop. Пакетная обработка данных в Hadoop. Рассматриваемые вопросы: - вычислительное ядро Hadoop; - концепция вычислений MapReduce; - архитектура MapReduce; - архитектура YARN.
4	Базы данных и Hadoop. Рассматриваемые вопросы: - SQL и NoSQL базы данных; - SQL и NoSQL инструменты в Hadoop; - введение в Hive, архитектура и модель данных в Hive; - диалект HiveQL, обработка данных в Hive.
5	Озеро данных. Рассматриваемые вопросы: - введение в понятие «озеро данных»; - приницпыи подходы к построению корпоративного озера данных; - конвейеры обработки данных; - принципы управления большими данными.
6	Введение в Apache Spark. Рассматриваемые вопросы: - введение в итеративную обработку больших данных на Apache Spark; - особенности обработки больших данных на Apache Spark, отличие от MapReduce; - Spark RDD API; - Spark Dataframe API;

№ п/п	Тематика лекционных занятий / краткое содержание
	- SQL-запросы на Spark; - GraphX и GraphFrames.
7	Введение в Apache Spark. Обработка данных в реальном времени. Рассматриваемые вопросы: - принципы обработки данных в реальном времени; - обзор возможностей в Spark Streaming API; - обработка данных в реальном времени с помощью Spark Streaming API.
8	Введение в Kafka. Распределенный брокер сообщений. Рассматриваемые вопросы: - область применения и примеры использования; - компоненты и архитектура Apache Kafka; - брокеры, поставщики и потребители данных, работа с сообщениями; - базовые операции Apache Kafka; - сценарии интеграции с Apache Kafka; - общие понятия Kafka Stream, работа с потоками; - обработка данных в в реальном времени из Kafka.
9	Развертывание инфраструктуры для сбора, хранения и обработки больших данных. Рассматриваемые вопросы: - развертывание Hadoop и его инструментов; - настройка Hive и HBase; - настройка Apache Spark; - развертывание Kafka.

4.2. Занятия семинарского типа.

Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	Сбор и хранение данных в распределенных файловых системах. В результате выполнения практической работы студент получает навык работы с HDFS для сбора и хранения данных.
2	Введение в MapReduce. В результате выполнения практической работы студент знакомится с концепцией MapReduce для обработки больших данных.
3	Обработка данных с помощью MapReduce. В результате выполнения практической работы студент получает навык обработки больших данных с использованием MapReduce.
4	Введение в Hive. В результате выполнения практической работы знакомится с диалектом HiveQL.
5	Обработка данных с помощью Hive. В результате выполнения практической работы студент получает навык обработки больших данных с использованием Hive.
6	Введение в Apache Spark. В результате выполнения практической работы студент знакомится с итеративной обработкой больших данных на Apache Spark.
7	Обработка данных с помощью Apache Spark. В результате выполнения практической работы студент получает навык обработки больших данных с использованием Apache Spark.

№ п/п	Тематика практических занятий/краткое содержание
8	Введение в Apache Kafka. В результате выполнения практической работы студент знакомится с распределенным брокером сообщений Apache Kafka.
9	Обработка данных с помощью Apache Kafka и Spark Streaming API. 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.

№ 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).

№	Библиографическое описание	Место доступа

п/п		
1	Чак, Л. Hadoop в действии / Л. Чак. — Москва : ДМК Пресс, 2012. — 424 с. — ISBN 978-5-94074-785-7. — Текст : электронный // Лань : электронно-библиотечная система.	URL: https://e.lanbook.com/book/39997 (дата обращения: 21.05.2023). — Режим доступа: для авториз. пользователей.
2	Макшанов, А. В. Большие данные. Big Data / А. В. Макшанов, А. Е. Журавлев, Л. Н. Тындыкарь. — 3-е изд., стер. — Санкт-Петербург : Лань, 2023. — 188 с. — ISBN 978-5-507-46866-9. — Текст : электронный // Лань : электронно-библиотечная система.	URL: https://e.lanbook.com/book/322664 (дата обращения: 21.05.2023). — Режим доступа: для авториз. пользователей.
3	Лебедев, А. С. Методы Big Data : учебно-методическое пособие / А. С. Лебедев, Ш. Г. Магомедов. — Москва : РТУ МИРЭА, 2021. — 91 с. — Текст : электронный // Лань : электронно-библиотечная система. Учебно-методическое издание	URL: https://e.lanbook.com/book/182452 (дата обращения: 21.05.2023). — Режим доступа: для авториз. пользователей.
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 // Лань : электронно-библиотечная система.	URL: https://e.lanbook.com/book/241001 (дата обращения: 21.05.2023). — Режим доступа: для авториз. пользователей.
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 // Лань : электронно-библиотечная система.	URL: https://e.lanbook.com/book/105840 (дата обращения: 21.05.2023). — Режим доступа: для авториз. пользователей.

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/)

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

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 ЖДСТУ Ю.О. Пазойский

Заведующий кафедрой ЦТУТП В.Е. Нутович

Председатель учебно-методической
комиссии Н.А. Андриянова