ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«БЕЛГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ»
(Н И У « Б е л Г У »)

УТВЕРЖДАЮ
Директор института инженерных и цифровых технологии

К.А. Польщиков
18.05.2022

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Технологии хранения и обработки больших объёмов данных
наименование дисциплины

Программа составлена в соответствии с требованиями ФГОС ВО по направлению подготовки

Направление подготовки 09.04.03 Прикладная информатика
Профиль подготовки Искусственный интеллект и наука о данных

Автор: доцент кафедры системного программирования СПбГУ, к.т.н. Серов Михаил Александрович
должность, ученая степень, ученое звание, инициалы и фамилия

Программа одобрена Кафедрой прикладной информатики информационных технологий

Протокол заседания кафедры от 06.04.2022 № 8
дата

Программа согласована Кафедрой прикладной информатики и информационных технологий

Протокол заседания кафедры от 06.04.2022 № 8
дата

Раздел 1. Характеристики учебных занятий

1.1. Цели и задачи учебных занятий

Дисциплина «Технологии хранения и обработки больших объёмов данных» представляет обучающимся комплекс знаний, умений и навыков для работы с технологиями, связанными с хранением, обработкой и анализом больших объёмов данных.

Цель изучения дисциплины: знакомство обучающихся с технологиями хранения и обработки больших данных, как получивших широкое распространение относительно недавно, таких как распределённые файловые системы и NoSQL СУБД, так и давно существующих на рынке, таких как реляционные СУБД.

1.2. Требования подготовленности обучающегося к освоению содержания учебных занятий (пререквизиты)

Максимальная эффективность программы будет обеспечена при условии, что обучающийся:

- понимает элементарных конструкций языка Python и способен писать несложные программы;
- знает основные алгоритмы и базовые структуры данных;
- знание реляционных СУБД и языка SQL.

1.3. Перечень результатов обучения (learning outcomes)

Способность применять средства промышленных реляционных СУБД, а также нереляционных систем для хранения и обработки больших объёмов данных

№	Наименование категории (группы) компетенций	Код и наименование компетенции	Код индикатора и индикатор достижения универсальной компетенции
	1	2	3
1	Профессиональные компетенции	ПКП-4-ИИР-ПК-1. Способен исследовать и разрабатывать архитектуры систем искусственного интеллекта для различных предметных областей на основе комплексов методов и инструментальных средств систем искусственного интеллекта	ПКП-4-ИИР-ПК-1.1. Исследует и разрабатывает архитектуры систем искусственного интеллекта для различных предметных областей ПКП-4-ИИР-ПК-1.2. Выбирает комплексы методов и инструментальных средств искусственного интеллекта для решения задач в зависимости от особенностей предметной области

1.4. Перечень и объём активных и интерактивных форм учебных занятий

Активные и интерактивные формы учебных занятий — семинары (10 ак. часов), лекции, предполагающие активное обсуждения материала с преподавателем (10 ак. часов).

Раздел 2. Организация, структура и содержание учебных занятий

2.1. Организация учебных занятий

2.1.1 Основной курс

Трудоёмкость, объёмы учебной работы и наполняемость групп обучающихся																	
Код модуля в составе дисциплины, практики и т.п.	Контактная работа обучающихся с преподавателем											Самостоятельная работа				Объём активных и	Трудоёмкость
	лекции	семинары	консультации	практические занятия	лабораторные работы	контрольные работы	коллоквиумы	текущий контроль	промежуточная аттестация	итоговая аттестация	под руководством	в присутствии	сам. раб. с использованием	текущий контроль	промежуточная аттестация (сам.раб.)	итоговая аттестация		
ОСНОВНАЯ ТРАЕКТОРИЯ																		
Форма обучения: очная																		
Семестр 4	16	14	2						2				74		36		25	4
	2-25	2-25	2-25						2-25				1-1		1-1			
ИТОГО	16	14	2						2				74		36		25	4

Виды, формы и сроки текущего контроля успеваемости и промежуточной аттестации						
Код модуля в составе дисциплины, практики и т.п.	Формы текущего контроля успеваемости		Виды промежуточной аттестации		Виды итоговой аттестации (только для программ итоговой аттестации и дополнительных образовательных программ)	
	Формы	Сроки	Виды	Сроки	Виды	Сроки
ОСНОВНАЯ ТРАЕКТОРИЯ						
Форма обучения: очная						
Семестр 4			экзамен, устно, традиционная форма	по графику промежуточной аттестации		

2.2. Структура и содержание учебных занятий

№ п/п	Наименование темы (раздела, части)	Вид учебных занятий	Количество часов
I	Модуль 1.1-3	лекции	4
		семинары	4
		по методическим материалам	18
II	Модуль 1.4-6	лекции	6
		семинары	4

		по методическим материалам	18
III	Модуль 1.6-8	лекции	3
		семинары	4
		по методическим материалам	18
IV	Модуль 1.8-10	лекции	3
		семинары	2
		по методическим материалам	18
	Промежуточная аттестация	самостоятельная работа	72
		консультации	2
		экзамен	6
Итого			

Модуль 1
1. Распределенные файловые системы.
2. Распределенная параллельная обработка данных технологией Map-Reduce.
3. Полнотекстовый поиск.
4. Статический ранг документов. Распределенные вычисления на графах.
5. NoSQL. Google Bigtable.
6. Согласованность в распределенных системах. Percolator.
7. Средства интеграции больших объёмов данных.
8. Создание ETL процесса: Case Study.
9. Поиск похожих документов.
10. Алгоритмы кластеризации.

Раздел 3. Обеспечение учебных занятий

3.1. Методическое обеспечение

3.1.1 Методические указания по освоению дисциплины

Литература: Тоби Сегаран «Программируем коллективный разум»; Jeffrey Ullman «Mining of Massive Datasets»; Christopher Manning «Introduction to Information Retrieval»

3.1.2 Методическое обеспечение самостоятельной работы

Статьи на соответствующие темы, опубликованные в журналах или сборниках трудов конференций; технические доклады и статьи сотрудников IT компаний, опубликованные на сайтах компаний

3.1.3 Методика проведения текущего контроля успеваемости и промежуточной аттестации и критерии оценивания

Оценка за экзамен ставится по следующим правилам: ответ на каждый вопрос билета и на дополнительные вопросы оценивается по шкале от 0 (нет ответа) до 10 (очень хороший ответ), далее оценка усредняется. Результат переводится в диапазон от 0 до 100. Далее применяется следующее правило выставления оценки:

Итоговый процент выполнения, %	Оценка СПбГУ при проведении экзамена	Оценка ECTS
90-100	отлично	A
80-89	хорошо	B

70-79	хорошо	C
61-69	удовлетворительно	D
50-60	удовлетворительно	E
менее 50	неудовлетворительно	F

3.1.4 Методические материалы для проведения текущего контроля успеваемости и промежуточной аттестации (контрольно-измерительные материалы, оценочные средства)

№	Код индикатора и индикатор достижения компетенции	Контрольно-измерительные материалы (КИМ) (тестовые вопросы, контрольные задания, кейсы и пр.)
	1	2
1	ПКП-4-ИИР-ПК-1.1. Исследует и разрабатывает архитектуры систем искусственного интеллекта для различных предметных областей	Контрольно-измерительные материалы устного экзамена
2	ПКП-4-ИИР-ПК-1.2. Выбирает комплексы методов и инструментальных средств искусственного интеллекта для решения задач в зависимости от особенностей предметной области	Контрольно-измерительные материалы устного экзамена

3.1.4.1 Формируемые дисциплиной компетенции

- ПКП-4-ИИР-ПК-1. Способен исследовать и разрабатывать архитектуры систем искусственного интеллекта для различных предметных областей на основе комплексов методов и инструментальных средств систем искусственного интеллекта

□ **Формируется дисциплиной.**

✓ **Развивается дисциплиной.**

□ **Полностью сформирована по результатам освоения дисциплины.**

Шкала оценивания: линейная, определяется долей ответов на вопросы, проверяющих данную компетенцию.

3.1.4.2 Контрольно-измерительные материалы (примеры)

Перечень примерных вопросов для экзамена:

1. Определение больших данных, ключевые характеристики. Примеры задач больших данных. Основные виды данных.
2. Дать краткую сравнительную характеристику инструментария ПО для анализа данных.
3. Охарактеризовать конструкции языка R Перечислить типы языка R, привести примеры.
4. Роль аналитика по данным (Data Scientist). Ключевые компетенции аналитика. Отличия BI от Data Science.
5. «Жизненный цикл» проекта по аналитике больших данных. Типовая архитектура проекта в области больших данных. Перечислить используемые технологии, указать степень вовлеченности каждой из технологий на каждом этапе работы над проектом. Перечислить основные роли исполнителей проекта.
6. Что такое Data Mining? Основные задачи и методы Data Mining. Этапы интеллектуального анализа данных. Методы интеллектуального анализа данных.

7. Что такое ИИ? Декатлон?
8. Роль гипотез в процессе познания. Какие факторы используются для уточнения гипотез?
9. Основные понятия статистики и дескриптивный анализ:
10. Шкалы измерений. Генеральная совокупность и выборка. Нормальное распределение. Уровень статистической достоверности.
11. Корреляция и регрессионный анализ. Коэффициент корреляции. Графическое представление. Постановка задачи регрессионного анализа.
12. Пояснить термин "Линейная регрессия". Привести примеры использования 12 регрессионного анализа.
13. Классификация и кластеризация – суть и назначение. Метрики. Постановка задачи кластеризации. Методы кластеризации на графах. Отличие от задачи классификации. Привести примеры использования алгоритмов кластеризации.
14. Парадигма Map Reduce. Описать принцип работы. Нарисовать схему. Перечислить слабые и сильные стороны. Обозначить области применимости. Привести примеры использования.
15. Визуализация. Дать определение визуализации. Показать важность визуализации в аналитике больших данных. Привести примеры и инструменты для визуализации.
16. Научные проблемы больших данных. Показать значимость проблем, актуальность, связь с областями математики и инженерии.
17. OLAP и OLTP системы. Разница.
18. Репликация и шардинг.
19. Требования ACID. CAP-теорема, BASE архитектура
20. NoSql. Классификация NoSql хранилищ. Их особенности. Примеры распределенных хранилищ.

Проверяемые компетенции: ОПК-2, ОПК-5, ПКП-4-ИИР-ПК-1

Критерии оценивания: совпадают с критериями оценивания ответа на экзамене.

Перечень примерных вопросов тестирования:

1. Формат Parquet считается
   a. неструктурированным
   b. полуструктурированным
   c. строковым
   d. колоночным (столбцовым) +
2. Для машинного обучения подходят данные
   a. Любых форматов в цифровом виде
   b. Числовые типа int
   c. Бинарные
   d. Предварительно подготовленные, очищенные от ошибок, пропусков и выбросов, а также нормализованные и представленные в виде числовых векторов +
3. Для полнотекстового интеллектуального поиска и аналитики по полуструктурированным данным в формате JSON отлично подходит СУБД
   a. HBase
   b. Cassandra
   c. Hive
   d. Elasticsearch +
4. Для распределенного глубокого машинного обучения (Deep Learning) больше подходит фреймворк
   a. TensorFlow
   b. Flask
   c. PyTorch +

d. Scikit-learn

5. Для реализации микросервисной архитектуры и интеграции разрозненных систем подходит
   a. Apache Kafka +
   b. Apache Spark
   c. Apache AirFlow
   d. Apache Hadoop
6. Apache NiFi используется для
   a. визуализации результатов аналитики
   b. эффективного хранения больших данных
   c. маршрутизации потоков Big Data и построения ETL-конвейеров +
   d. оптимизации SQL-запросов к DWH
7. Повысить производительность Apache Kafka можно с помощью:
   a. Увеличения плотности разделов на каждом брокере
   b. Повышения коэффициента репликации
   c. Увеличения размера сообщений
   d. Замены HDD-дисков на SSD +
8. Автоматизировать запуск пакетных задач в рамках конвейера обработки больших данных по расписанию можно с помощью
   a. Apache Hive
   b. Apache Hadoop
   c. Apache Kafka
   d. Apache AirFlow +
9. Выберите технологию потоковой обработки событий в режиме реального времени
   a. Spark Streaming
   b. Apache Kafka +
   c. Apache Hadoop
   d. MapReduce

3.1.5 Методические материалы для оценки обучающимися содержания и качества учебного процесса

Для оценки обучающимися содержания и качества учебного процесса применяется анкетирование в соответствии с методикой и графиком, утвержденными в установленном порядке.

3.2. Кадровое обеспечение

3.2.1 Образование и (или) квалификация штатных преподавателей и иных лиц, допущенных к проведению учебных занятий

К чтению лекций должны привлекаться преподаватели, имеющие ученую степень доктора или кандидата наук (в том числе степень PhD, прошедшую установленную процедуру признания и установления эквивалентности) и/или ученое звание профессора или доцента.

3.2.2 Обеспечение учебно-вспомогательным и (или) иным персоналом

Учебно-вспомогательный и инженерно-технический персонал должен иметь соответствующее образование и обладать навыками организации работы с пользовательскими программными продуктами в локальной сети компьютерного класса и в Интернете.

3.3. Материально-техническое обеспечение

3.3.1 Характеристики аудиторий (помещений, мест) для проведения занятий

В аудиториях, где проводятся занятия, необходимо наличие поверхностей для сидения (стульев), поверхностей для письма (столов или откидных столиков), досок и средств письма на них.

3.3.2 Характеристики аудиторного оборудования, в том числе неспециализированного компьютерного оборудования и программного обеспечения общего пользования

Аудитории для проведения занятий должны быть оснащены проекционной техникой и компьютером. Желательно наличие выхода в интернет.

3.3.3 Характеристики специализированного оборудования

Специальных требований нет

3.3.4 Характеристики специализированного программного обеспечения

Специальных требований нет

3.3.5 Перечень и объёмы требуемых расходных материалов

Специальных требований нет

3.4. Информационное обеспечение

Не требуется.

3.4.1 Список литературы

1. Тарасов, С. В. СУБД для программиста: базы данных изнутри / С. В. Тарасов. - Москва : СОЛОН-Пресс, 2020. - 320 с. - ISBN 978-2-7466-7383-0. – ЭР СПбГУ: https://proxy.library.spbu.ru/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat07918a&AN=spsu.znanium369884&lang=ru&site=eds-live&scope=site
2. Мартишин, С. А. Проектирование и реализация баз данных в СУБД MySQL с использованием MySQL Workbench. Методы и средства проектирования информационных систем и технологий. Инструментальные средства информационных систем : учебное пособие / С.А. Мартишин, В.Л. Симонов, М.В. Храпченко. — Москва : ФОРУМ : ИНФРА-М, 2021. — 160 с. - ISBN 978-5-8199-0811-2. – ЭР СПбГУ: https://proxy.library.spbu.ru/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat07918a&AN=spsu.znanium365080&lang=ru&site=eds-live&scope=site
3. Дадян, Э. Г. Данные: хранение и обработка : учебник / Э. Г. Дадян. — Москва : ИНФРА-М, 2021. — 205 с. — (Высшее образование: Бакалавриат). - ISBN 978-5-16-016447-2. – ЭР СПбГУ: https://proxy.library.spbu.ru/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat07918a&AN=spsu.znanium360938&lang=ru&site=eds-live&scope=site
4. Макшанов А. В., Журавлев А. Е., Тындыкарь Л. Н. Большие Данные. Big Data. Издательство “Лань”; 2021. 188 с. – ЭР СПбГУ: https://proxy.library.spbu.ru/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat07918a&AN=spsu.lanbook165835&lang=ru&site=eds-live&scope=site
5. Эрик Редмонд, Джим. Р. Уилсон. Семь Баз Данных За Семь Недель. Введение в Современные Базы Данных и Идеологию NoSQL. Издательство “ДМК Пресс”; 2013. 384 с. – ЭР СПбГУ: https://proxy.library.spbu.ru/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=cat07918a&AN=spsu.lanbook58690&lang=ru&site=eds-live&scope=site

3.4.2 Перечень иных информационных источников, в том числе современных профессиональных баз данных и информационных справочных систем

Электронные ресурсы Научной библиотеки им. М. Горького СПбГУ

- Сайт Научной библиотеки им. М. Горького СПбГУ: https://library.spbu.ru/ru/
- Электронный каталог Научной библиотеки им. М. Горького СПбГУ: http://old.library.spbu.ru/cgi-bin/irbis64r/cgiirbis_64.exe?C21COM=F&I21DBN=IBIS&P21DBN=IBIS
- Перечень электронных ресурсов, находящихся в доступе СПбГУ: http://cufts.library.spbu.ru/CRDB/SPBGU/
- Перечень ЭБС, на платформах которых представлены российские учебники, находящиеся в доступе СПбГУ: http://cufts.library.spbu.ru/CRDB/SPBGU/browse?resource_type=8&name=rures .
- Перечень ресурсов и баз данных по тематике Математика http://cufts.library.spbu.ru/CRDB/SPBGU/browse?subject=1
- Перечень ресурсов и баз данных по тематике Информатика http://cufts.library.spbu.ru/CRDB/SPBGU/browse?subject=93

Раздел 4. Разработчики программы

Серов Михаил Александрович, к.т.н., доцент, кафедра системного программирования СПбГУ, m.serov@spbu.ru, +7 (962) 2951889