**Санкт-Петербургский государственный университет**

**Р А Б О Ч А Я П Р О Г Р А М М А**

**УЧЕБНОЙ ДИСЦИПЛИНЫ**

Технологии хранения и обработки больших объёмов данных

Big Data Storage and Processing Techniques

**Язык(и) обучения**

русский

Трудоемкость в зачетных единицах: 3

Регистрационный номер рабочей программы: 027268

Санкт-Петербург

2020

**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)**

Программа курса предназначена обучающимся 1 курса магистратуры и рассчитана на обучающихся, изучавших основы математических дисциплин в объеме четырех курсов бакалавриата.

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

Курс способствует формированию следующих компетенций:

* ПКА-1 – способен находить, формулировать и решать актуальные проблемы фундаментальной и прикладной информатики и информационных технологий;
* ПКП-1 – способен демонстрировать фундаментальные знания математических и естественных наук, программирования и информационных технологий;
* ПКП-7 – способен использовать знания направлений развития компьютеров с традиционной (нетрадиционной) архитектурой; современных системных программных средств: операционных систем, операционных и сетевых оболочек, сервисных программ; тенденции развития функций и архитектур проблемно-ориентированных программных систем и комплексов в профессиональной деятельности;
* ПКП-8 – способен использовать основные концептуальные положения функционального, логического, объектно-ориентированного и визуального направлений программирования, методы, способы и средства разработки программ в рамках этих направлений.

**1.4. Перечень и объём активных и интерактивных форм учебных занятий**

Активные и интерактивные формы учебных занятий — семинары (15 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 интерактивных  форм учебных занятий | Трудоёмкость |
| лекции | семинары | консультации | практические  занятия | лабораторные работы | контрольные работы | коллоквиумы | текущий контроль | промежуточная  аттестация | итоговая аттестация | под руководством преподавателя | в присутствии  преподавателя | сам. раб. с использованием  методических материалов | текущий контроль (сам.раб.) | промежуточная аттестация (сам.раб.) | итоговая аттестация  (сам.раб.) |
| ОСНОВНАЯ ТРАЕКТОРИЯ | | | | | | | | | | | | | | | | | | |
| Форма обучения: очная | | | | | | | | | | | | | | | | | | |
| Семестр 2 | 15 | 15 | 2 |  |  |  |  |  | 2 |  |  |  | 44 |  | 30 |  | 25 | 3 |
|  | 2-100 | 10-25 | 2-100 |  |  |  |  |  | 2-100 |  |  |  | 1-1 |  | 1-1 |  |  |  |
| ИТОГО | 15 | 15 | 2 |  |  |  |  |  | 2 |  |  |  | 44 |  | 30 |  |  | 3 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 2 |  |  | экзамен, устно, традиционная форма | по графику промежуточной аттестации |  |  |

**2.2. Структура и содержание учебных занятий**

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 | А |
| 80-89 | хорошо | В |
| 70-79 | хорошо | С |
| 61-69 | удовлетворительно | D |
| 50-60 | удовлетворительно | Е |
| менее 50 | неудовлетворительно | F |

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

Формируемые дисциплиной компетенции:

* ПКА-1 – способен находить, формулировать и решать актуальные проблемы фундаментальной и прикладной информатики и информационных технологий;
* ПКП-1 – способен демонстрировать фундаментальные знания математических и естественных наук, программирования и информационных технологий;
* ПКП-7 – способен использовать знания направлений развития компьютеров с традиционной (нетрадиционной) архитектурой; современных системных программных средств: операционных систем, операционных и сетевых оболочек, сервисных программ; тенденции развития функций и архитектур проблемно-ориентированных программных систем и комплексов в профессиональной деятельности;
* ПКП-8 – способен использовать основные концептуальные положения функционального, логического, объектно-ориентированного и визуального направлений программирования, методы, способы и средства разработки программ в рамках этих направлений.

Шкала оценивания: линейная, определяется долей успешно выполненных заданий, проверяющих компетенцию.

*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*: ПКА-1, ПКП-1, ПКП-7, ПКП-8

*Критерии оценивания*: совпадают с критериями оценивания ответа на экзамене.

**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 преподаванию дисциплины могут быть допущены преподаватели, имеющие диплом о высшем образовании по соответствующему направлению.

**3.2.2 Обеспечение учебно-вспомогательным и (или) иным персоналом**

Специальных требований нет.

**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 обязательной литературы**

Не требуется.

**3.4.2 Список дополнительной литературы**

Не требуется.

**3.4.3 Перечень иных информационных источников**

Специальных требований нет

**Раздел 4. Разработчики программы**

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