Санкт-Петербургский государственный университет

Р А Б О Ч А Я П Р О Г Р А М М А
УЧЕБНОЙ ДИСЦИПЛИНЫ

Теория и практика больших данных

Big Data

Язык(и) обучения

русский

Трудоемкость в зачетных единицах: 3

Регистрационный номер рабочей программы: 053716

Раздел 1. Характеристики учебных занятий

1.1. Цели и задачи учебных занятий

Изучение основных алгоритмов, подходов и актуальных программных средств обработки и анализа «больших данных».

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

Знание базовых алгоритмов, структур данных и их свойств, знание какого-либо современного языка программирования высокого уровня, знание реляционных баз данных и языка SQL.

1.3. Перечень результатов обучения (learning outcomes)

В результате освоения дисциплины обучающийся должен иметь общее представление о современных методах обработки больших объемов данных, понимать принципы построения соответствующих алгоритмов и систем хранения, ориентироваться в экосистемах Hadoop и Spark, уметь писать приложения под указанные платформы.

1.4. Перечень и объём активных и интерактивных форм учебных занятий

Практические занятия, задания на реальных данных, работа с кластером Spark.

Раздел 2. Организация, структура и содержание учебных занятий

2.1. Организация учебных занятий

2.1.1 профиль Технологии баз данных

Трудоёмкость, объёмы учебной работы и наполняемость групп обучающихся																	
Код модуля в составе дисциплины, практики и т.п.	Контактная работа обучающихся с преподавателем											Самостоятельная работа				Интерактивных	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	16		2	16					2				48		24		16	3
	2-25		2-25	10-25					2-100				1-1		1-1			
ИТОГО	16		2	16					2				48		24			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 обученияочная						
Семестр 2			экзамен, устно, домашние задания	по графику промежуточной аттестации		

2.2. Структура и содержание учебных занятий

№ п/п	Наименование темы (раздела, части)	Вид учебных занятий	Количество часов
1	Введение	лекции	2
		практические занятия	0
		по методическим материалам	6
2	Язык Python для анализа данных	лекции	5
		практические занятия	6
		по методическим материалам	14
3	Инструментарий	лекции	5
		практические занятия	4
		по методическим материалам	14
4	Apache Spark	лекции	4
		практические занятия	6
		по методическим материалам	14

Содержание учебных занятий

1. Введение. Мотивация, «эпоха больших данных». Масштабирования информационных систем, почему не работают классические подходы к хранению и обработке данных. «Теорема» CAP. Распределенные вычисления и обработка данных.

2. Язык Python для анализа данных. Введение в язык,работа в Jupyter Notebook. NumPy, SciPy, основные концепции. Matplotlib и seaborn, построение графиков и визуализация данных. Scikit-learn, базовый API, процесс работы с данными. Pandas, операции, преобразование данных, визуализация, производительность, работы в связке с базами данных. XGBoost, Vowpal Wabbit.

3. Инструментарий. JVM и Scala. Управление зависимостями. Maven/Gradle/SBT. REPL. Система типов, структуры данных, операции с данными. Параллельные приложения, работа с потоками, JMM. Примитивы синхронизации, основные паттерны. Lock-Free структуры данных. Модель акторов, Akka. Другие парадигмы распределенных вычислений.

4. Apache Spark. Концепция RDD и исторические проблемы, ленивость операций. DataFrames, SparkSQL, PySpark. Запуск приложений, работа с кластером, мониторинг. Производительность, память, локальность данных, жизненный цикл программы, кеширование. Высокоуровневая работа с данными - Интеграция с другими источниками данных. Spark Streaming, мотивация и особенности. Apache Kafka. Машинное обучение с использованием Spark. Mlib, GraphX.

Раздел 3. Обеспечение учебных занятий

3.1. Методическое обеспечение

3.1.1 Методические указания по освоению дисциплины

Список используемой обязательной литературы.

3.1.2 Методическое обеспечение самостоятельной работы

Практические задания, электронные дополнительные материалы.

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

Основными документами, регламентирующими порядок организации и проведения текущего контроля успеваемости, промежуточной аттестации обучающихся являются: "Правила обучения по основным образовательным программам высшего и среднего образования СПбГУ".
На первом занятии преподаватель доводит до сведения обучающихся график (сроки) текущего контроля их самостоятельной работы и критерии оценки знаний при текущем контроле успеваемости, а также сроки и условия заключительной (промежуточной) аттестации.
На экзамене обучающийся должен предоставить два выполненных домашних задания, решить практическую задачу по пройденному материалу, ответить на вопрос из экзаменационного билета. Всего к экзамену предлагается список из 10 билетов. Оценка «отлично» ставится если обучающийся без существенных замечаний выполнил два домашних задания, практическое задание и ответил на вопрос. Если один из пунктов не выполнен, то ставится оценка «хорошо», если два - «удовлетворительно», в противном случае - «неудовлетворительно».
Реализацию непрерывного контроля знаний согласно графику, преподаватель осуществляет за счет часов, предусмотренных нормами времени на проверку различного рода письменных работ, проведение консультаций и пр.
Преподаватель имеет право изменять структуру и количество модулей дисциплины и разделов в них, в зависимости от изменения нормативной базы и количество точек контроля знаний слушателей за период обучения. Однако при этом необходимо обеспечить соответствие затрат учебного времени на самостоятельную работу слушателей установленным нормам затрат времени на эти виды контроля, а также бюджету времени, предусмотренного учебным планом на данную дисциплину

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

Два практических задания, решение которых нужно предоставить в экзамену, одно практическое задание на экзамене, список из 10 вопросов .

Пример практического задания:

> Напишите две программы: первая генерирует бинарный файл (min 2Гб), состоящий из случайных 32-рязрядных беззнаковых целых чисел (big endian).
> Вторая считает сумму этих чисел (с применением длинной арифметики), находит минимальное и

максимальное число.

Реализуйте две версии -
а. Простое последовательное чтение
б. Многопоточная + memory-mapped files. Сравните время работы.

Примерный список вопросов:

1. Большие данные и `NoSQL`. Основные причины возникновения явления.
2. "Теорема" `CAP`.
3. Экосистема `Hadoop`, `Apache Hadoop YARN` - зачем нужен и какие задачи решает. Архитектура `HDFS`.
4. Модель памяти `JVM`, основные идеи. Отношение `happens-before`. Ключевое слово `volatile`,
5. Примитивы синхронизации в `JVM` (`java.util.concurrent`)
6. Модель акторов, основные идеи, достоинства и недостатки. Примеры на `Akka`.
7. Работа с потоками и процессами из `Python`. `GIL`. Модуль `multiprocessing`, способы реализации (spawn, fork, forkserver)
8. `Apache Spark` - общее устройство. Работа с `HDFS`, data locality.
9. `RDD`, достоинства и недостатки.
10. `Dataset API`, `Spark SQL` & `DataFrames` - мотивация, основные идеи.

3.1.5 Методические материалы для оценки обучающимися содержания и качества учебного процесса

Просим Вас заполнить анкету-отзыв по прочитанной дисциплине.

Обобщенные данные анкет будут использованы для ее

совершенствования. По каждому вопросу проставьте соответствующие оценки по шкале от 1 до 10 баллов (обведите выбранный Вами балл). В случае необходимости впишите свои комментарии.

1. Насколько Вы удовлетворены содержанием дисциплины в

целом? 1 2 3 4 5 6 7 8 9 10

Комментарий_____

2. Насколько Вы удовлетворены общим стилем преподавания?

1 2 3 4 5 6 7 8 9 10

Комментарий_____

3. Как Вы оцениваете качество подготовки предложенных

методических материалов? 1 2 3 4 5 6 7 8 9 10

Комментарий_____

4. Насколько Вы удовлетворены использованием

преподавателями активных методов обучения? 1 2 3 4 5 6 7 8 9 10

Комментарий_____

5. Какой из модулей (разделов) дисциплины Вы считаете

наиболее полезным, ценным с точки зрения дальнейшего обучения и/или применения в последующей практической деятельности?

Комментарий_____

6. Что бы Вы предложили изменить в методическом и

содержательном плане для совершенствования преподавания данной

дисциплины?

Комментарий_____

СПАСИБО!

3.2. Кадровое обеспечение

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

К преподаванию допускаются преподаватели, владеющие соответствующим материалом и имеющим практический опыт применения изучаемых технологий.

3.2.2 Обеспечение учебно-вспомогательным и (или) иным персоналом

Отсутствуют.

3.3. Материально-техническое обеспечение

3.3.1 Характеристики аудиторий (помещений, мест) для проведения занятий

Компьютерный класс с доступом в интернет.

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

Проектор.

3.3.3 Характеристики специализированного оборудования

Нет специальных требований.

3.3.4 Характеристики специализированного программного обеспечения

Доступ к кластеру Hadoop. Доступ к кластеру Spark. Компьютеры с VirtualBox.

3.3.5 Перечень и объёмы требуемых расходных материалов

Нет специальных требований.

3.4. Информационное обеспечение

3.4.1 Список обязательной литературы

1. https://docs.oracle.com/javase/specs/jls/se7/html/jls-17.html
2. https://spark.apache.org/documentation.html
3. https://akka.io/docs
4. https://docs.python.org/3/library/multiprocessing.html

3.4.2 Список дополнительной литературы

1. http://greenteapress.com/semaphores/LittleBookOfSemaphores.pdf

2. Холден Карау, «Энди Конвински, Патрик Венделл, Матей Захария, «Изучаем Spark. Молниеносный анализ данных», ДМК Пресс, 2015
3. Чак Лэм, «Hadoop в действии», ДМК Пресс , 2012
4. Donald Miner, Adam Shook, «MapReduce Design Patterns: Building Effective Algorithms

and Analytics for Hadoop and Other Systems», O'Reilly Media; 1 edition, 2012
5. Alex Holmes, «Hadoop in Practice», Manning, Second Edition, 2012

3.4.3 Перечень иных информационных источников

Нет специальных требований.

Раздел 4. Разработчики программы

Мишенин Алексей Николаевич, a.mishenin@spbu.ru