Санкт-Петербургский государственный университет

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

Hadoop

Hadoop

Язык(и) обучения

русский

Трудоёмкость (границы трудоёмкости) в зачетных единицах: 2

Регистрационный номер рабочей программы: _____

Санкт-Петербург

2017

Раздел 1. Характеристики учебных занятий

1.1. Цели и задачи учебных занятий

Изучение архитектуры, сценариев использования, алгоритмов, методов диагностики и отладки программных инструментов из экосистемы «Hadoop».

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 разбираться в архитектуре программных средств из экосистемы «Hadoop» , должен уметь применять их на практике, разрабатывать программные решения.

1.4. Перечень и объём активных и интерактивных форм учебных занятий

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

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																		
Период обучения (модуль)	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																		
очная форма обучения																		
Семестр 3		12		16				4	2				26	4	8		16	2
		2-15		2-15				2-15	2-15				1-1	1-1	1-1			
ИТОГО		**12**		**16**				**4**	**2**				**26**	**4**	**8**			**2**

Формы текущего контроля успеваемости, виды промежуточной и итоговой аттестации			
Период обучения (модуль)	Формы текущего контроля успеваемости	Виды промежуточной аттестации	Виды итоговой аттестации (только для программ итоговой аттестации и дополнительных образовательных программ)
ОСНОВНАЯ ТРАЕКТОРИЯ			
очная форма обучения			
Семестр 3	доклад, три домашних задания	зачет	

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	Архитектура больших распределенных систем. Распределенные файловые системы.	лекции	2
		практические занятия	6
		по методическим материалам	14
3	Hadoop	лекции	7
		практические занятия	4
		по методическим материалам	14
4	Разработка распределенных приложений	лекции	4
		практические занятия	6
		по методическим материалам	14

Содержание учебных занятий

1. Введение. Мотивация, «эпоха больших данных». Горизонтальное/вертикальное масштабирование. История — MapReduce, BigTable. Примеры задач, современное состояние. Дистрибутивы Hadoop. Установка тестового окружения (виртуальные машины/Docker). Текущие положение в области программных средств для обработки больших данных и Hadoop.

2. Архитектура больших распределенных систем. Распределенные файловые системы. Обзор архитектуры и парадигма вычислений, HDFS (обзор), основные требования к надежности, Data locality, Fault Tolerance, Hadoop первой версии — проблемы, YARN.

3. Hadoop. JVM — особенности в плане решаемых задач. Парадигма MapReduce. HDFS - репликация, надежность, настройка, кеширование, командная строка. YARN - основные концепции, контейнеры и планировщики. Mapper, Reducer, Combiner, Partitioners, InputFormat, OutputFormat. Типы данных, механизм сериализации. Жизненый цикл задачи. Кеширование данных. Счетчики. Выбор числа map- и reduce- задач. Отладка, логирование, контроль.

4. Разработка распределенных приложений. Особенности реализации NoSQL баз данных. HBase. Cassandra. Hive. Алгоритмы для MapReduce. Spark - DataFrames, SparkSQL. 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.

Для получения зачета обучающийся должен предоставить три выполненных домашних задания и сделать доклад в течение семестра.

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)

Примеры домашних заданий:

1. Нужно написать скрипт, который скачивает данные последних президентских выборов для всех избирательных участков. С помощью Spark или Hadoop посчитать:
   - явку (%) по всем регионам, результат отсортировать по убыванию
   - выбрать произвольного кандидата и найти тот избирательный участок, на котором он получил наибольший результат (учитывать участки на которых проголосовало больше 300 человек)
   - найти регион, где разница между ТИК с наибольшей явкой и наименьшей максимальна
   - посчитать дисперсию по явке для каждого региона (учитывать УИК)
2. Используя VK Streaming API, Kafka и Spark Streaming реализуйте систему мониторинга событий в социальной сети VK.
3. С помощью jpcap и Spark Streaming реализуйте систему контроля трафика на локальной машине.

Тема доклада выбирается обучающимся самостоятельно, после согласования с преподавателем. Тема может быть может быть выбрана по следующим критериям:

1. Подробный доклад о алгоритмическом устройстве одной из рассматриваемых в курсе технологий. Например: «Генератор кода в Tungsten», «Система компрессии данных в Apache Hadoop», «Устройство хранилища данных в HBase».

2. Разбор научной статьи по теме курса. Статья должна быть опубликована в течение последнего года.

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/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://hadoop.apache.org/docs/current/

2. https://spark.apache.org/documentation.html

3. https://hbase.apache.org/book.html

4. https://hive.apache.org/

3.4.2 Список дополнительной литературы

1. Холден Карау, «Энди Конвински, Патрик Венделл, Матей Захария, «Изучаем Spark. Молниеносный анализ данных», ДМК Пресс, 2015
2. Чак Лэм, «Hadoop в действии», ДМК Пресс , 2012
3. Donald Miner, Adam Shook, «MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems», O'Reilly Media; 1 edition, 2012
4. 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