Федеральное государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт (национальный исследовательский университет)»

УТВЕРЖДЕНО
Начальник учебного управления

И.Р. Гарайшина

Рабочая программа дисциплины (модуля)

по дисциплине: Алгоритмы обработки больших данных
по направлению: Программная инженерия
профиль подготовки: Разработка программно-информационных систем
высшая школа программной инженерии
высшая школа программной инженерии МФТИ - Яндекс
курс: 3
квалификация: бакалавр

Семестр, формы промежуточной аттестации: 6 (весенний) - Дифференцированный зачет

Аудиторных часов: 60 всего, в том числе:
лекции: 30 час.
семинары: 30 час.
лабораторные занятия: 0 час.

Самостоятельная работа: 48 час.

Всего часов: 108, всего зач. ед.: 3

Программу составил: А.В. Малеев, директор

Программа обсуждена на заседании высшей школы программной инженерии МФТИ - Яндекс 08.06.2022

Аннотация

Дисциплина «Алгоритмы обработки больших данных» обеспечивает фундаментальное приобретение знаний и умений в области информатики и основ программирования. Прогрмамма имеет чётко выраженную прикладную направленность полученные знани я и навыки будут использованы в дальнейшем при изучении профессиональных и специальных дисциплин компьютерного цикла.

1. Цели и задачи

Цель дисциплины

- формирование у студентов знаний о принципах и инструментарии пакетной и потоковой обработки больших объёмов данных.

Задачи дисциплины

- систематизации навыков проектирования архитектур, применения специализированных инструментов и разработки программных систем для работы с большими объемами данных;
- формирование понимания внутреннего устройства, механики работы, области применимости существующих решений.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.2 Находит, критически анализирует и выбирает информацию, необходимую для решения поставленной задачи
ОПК-1 Способен применять естественнонаучные и общеинженерные знания, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности	ОПК-1.1 Способен анализировать поставленную задачу, намечать пути ее решения
	ОПК-1.3 Способен определять границы применимости полученных результатов
ОПК-7 Способен применять в практической деятельности основные концепции, принципы, теории и факты, связанные с информатикой	ОПК-7.1 Обладает навыками создания и выполнения тестовых сценариев для выявления ошибок в программном обеспечении
	ОПК-7.2 Понимает принципы работы баз данных и умеет проектировать структуру данных для эффективного хранения информации
ОПК-8 Способен осуществлять поиск, хранение, обработку и анализ информации из различных источников и баз данных, представлять ее в требуемом формате с использованием информационных, компьютерных и сетевых технологий	ОПК-8.1 Понимает принципы, по которым работают базы данных, и умеет создавать структуру данных, оптимизированную для эффективного хранения и обработки информации
	ОПК-8.2 Умеет примененять технологии машинного обучения в различных прикладных областях
	ОПК-8.3 Умеет оптимизировать и проводить рефакторинг существующего кода для улучшения производительности и поддержки
ПК-1 Способен самостоятельно или в качестве члена малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-1.2 Способен проводить научные исследования самостоятельно или в качестве члена малого научного коллектива

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- типы хранилищ больших объёмов данных;
- подходы к потоковой и пакетной обработке данных;
- принципы трансляции высокоуровневых языков программирования (SQL-подобных и функциональных) в последовательность задач на Hadoop кластере;
- основные определения и понятия потоковой обработки больших данных;
- основы обработки данных в реальном времени.

уметь:

- пользоваться распределенной файловой системой;
- запускать задачи на Hadoop кластере;
- писать задачи для запуска на Hadoop кластере с помощью нативного Java-интерфейса;
- писать задачи для запуска на Hadoop кластере с помощью любого другого языка программирования (с помощью инструментария Hadoop streaming);
- решать задачи статистики, задачи поиска и индексации, задачи машинного обучения на Hadoop кластере.

владеть:

- навыками работы с большими объемами данных и кругозором в выборе архитектурного решения поставленной задачи.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Распределённые файловые системы (GFS, HDFS)	2	2		4
2	Парадигма MapReduce	6	6		7
3	Управление ресурсами Hadoop-кластера. YARN	2	2		7
4	SQL over BigData: Apache Drill, Cloudera Impala, Presto, Hive	4	4		8
5	Технологии обработки данных в распределенной оперативной памяти. Apache Spark	6	6		7
6	Обработка данных в реальном времени. Kafka, Spark Streaming	4	4		8
7	BigData NoSQL, Key-value базы данных	6	6		7
Итого часов		30	30		48
Подготовка к экзамену		0 час.			
Общая трудоёмкость		108 час., 3 зач.ед.			

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 6 (Весенний)

1. Распределённые файловые системы (GFS, HDFS)

Распределённые файловые системы (GFS, HDFS). Её составляющие. Их достоинства, недостатки и сфера применения. Чтение и запись в HDFS. HDFS APIs: Web, shell, Java.

2. Парадигма MapReduce

Парадигма MapReduce. Основная идея, формальное описание. Обзор реализаций. Виды и классификация многопроцессорных вычислительных систем. Hadoop. Схема его работы, роли серверов в Hadoop-кластере. API для работы с Hadoop (Native Java API vs. Streaming), примеры.
MapReduce, продолжение. Типы Join'ов и их реализации в парадигме MR. Паттерны проектирования MR (pairs, stripes, составные ключи).

3. Управление ресурсами Hadoop-кластера. YARN

Hadoop MRv1 vs. YARN. Нововведения в последних версиях Hadoop. Планировщик задач в YARN. Apache Slider.

4. SQL over BigData: Apache Drill, Cloudera Impala, Presto, Hive

SQL over BigData: Apache Drill, Cloudera Impala, Presto, Hive. Повторение SQL. HiveQL vs. SQL. Виды таблиц в Hive, типы данных, трансляция Hive-запросов в MapReduce-задачи.
Аналитические функции в Hive. Расширения Hive: Streaming, User defined functions. Оптимизация запросов в Hive.

5. Технологии обработки данных в распределенной оперативной памяти. Apache Spark

Spark RDD vs Spark Dataframes
Spark SQL
Spark GraphFrames

6. Обработка данных в реальном времени. Kafka, Spark Streaming

Обработка данных в реальном времени. Spark Streaming.
Распределённая очередь Apache Kafka. Kafka streams.

7. BigData NoSQL, Key-value базы данных

HBase. NoSQL подходы к реализации распределенных баз данных, key-value хранилища. Основные компоненты BigTable-подобных систем и их назначение, отличие от реляционных БД. Чтение, запись и хранение данных в HBase. Minor- и major-компактификация. Надёжность и отказоустойчивость в HBase.
Cassandra. Основные особенности. Чтение и запись данных. Отказоустойчивость. Примеры применения HBase и Cassandra.
Отличие архитектуры HBase от Cassandra.

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

компьютер, мультимедийное оборудование (проектор).

6.Перечень рекомендуемой литературы

Основная литература

1. Адаптивная фильтрация сигналов: теория и алгоритмы, Электронная версия печатной публикации / В. И. Джиган. — Москва, Техносфера, 2013

-

Дополнительная литература

Литература, рекомендованная для самостоятельного изучения: Guller, M. (2015). Big Data Analytics with Spark : A Practitioner’s Guide to Using Spark for Large Scale Data Analysis. [Berkeley, CA]: Apress. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&site=eds-live&db=edsebk&AN=1174460

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

Не используются

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

не требуется.

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Программа курса в разделе «самостоятельная работа» обозначает минимально необходимое время для работы студента над темой. Самостоятельная работа включает в себя:

- проработку учебного материала (по конспектам лекций, учебной и научной литературе), подготовку ответов на вопросы, предназначенных для самостоятельного изучения, доказательство отдельных утверждений, свойств,
- выполнение домашних заданий.

ПРИЛОЖЕНИЕ

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Программная инженерия
профиль подготовки: Разработка программно-информационных систем
высшая школа программной инженерии МФТИ - Яндекс
высшая школа программной инженерии
курс: 3
квалификация: бакалавр

Семестр, формы промежуточной аттестации: 6 (весенний) - Дифференцированный зачет

Разработчик: А.В. Малеев, директор

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.2 Находит, критически анализирует и выбирает информацию, необходимую для решения поставленной задачи
ОПК-1 Способен применять естественнонаучные и общеинженерные знания, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности	ОПК-1.1 Способен анализировать поставленную задачу, намечать пути ее решения
	ОПК-1.3 Способен определять границы применимости полученных результатов
ОПК-7 Способен применять в практической деятельности основные концепции, принципы, теории и факты, связанные с информатикой	ОПК-7.1 Обладает навыками создания и выполнения тестовых сценариев для выявления ошибок в программном обеспечении
	ОПК-7.2 Понимает принципы работы баз данных и умеет проектировать структуру данных для эффективного хранения информации
ОПК-8 Способен осуществлять поиск, хранение, обработку и анализ информации из различных источников и баз данных, представлять ее в требуемом формате с использованием информационных, компьютерных и сетевых технологий	ОПК-8.1 Понимает принципы, по которым работают базы данных, и умеет создавать структуру данных, оптимизированную для эффективного хранения и обработки информации
	ОПК-8.2 Умеет примененять технологии машинного обучения в различных прикладных областях
	ОПК-8.3 Умеет оптимизировать и проводить рефакторинг существующего кода для улучшения производительности и поддержки
ПК-1 Способен самостоятельно или в качестве члена малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-1.2 Способен проводить научные исследования самостоятельно или в качестве члена малого научного коллектива

2. Показатели оценивания компетенций

В результате изучения дисциплины «Алгоритмы обработки больших данных» обучающийся должен:

знать:

- типы хранилищ больших объёмов данных;
- подходы к потоковой и пакетной обработке данных;
- принципы трансляции высокоуровневых языков программирования (SQL-подобных и функциональных) в последовательность задач на Hadoop кластере;
- основные определения и понятия потоковой обработки больших данных;
- основы обработки данных в реальном времени.

уметь:

- пользоваться распределенной файловой системой;
- запускать задачи на Hadoop кластере;
- писать задачи для запуска на Hadoop кластере с помощью нативного Java-интерфейса;
- писать задачи для запуска на Hadoop кластере с помощью любого другого языка программирования (с помощью инструментария Hadoop streaming);
- решать задачи статистики, задачи поиска и индексации, задачи машинного обучения на Hadoop кластере.

владеть:

- навыками работы с большими объемами данных и кругозором в выборе архитектурного решения поставленной задачи.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

С целью контроля освоения обучающимися учебного материала проводится устный опрос в начале занятия по теме прошлого занятия.

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

Примерный перечень вопросов для промежуточной аттестации:

1. Опишите устройство HDFS, основные идеи,
2. Опишите схему чтения и записи в HDFS.
3. Опишите основные положения идеи MapReduce, примеры применения.
4. Проблемы распределенных вычислений
5. Опишите подходы к реализации распределенных баз данных
6. Отличие архитектуры HBase от Cassandra.
7. Назовите виды таблиц в Hive
8. Паттерны проектирования MR (pairs, stripes, составные ключи),
9. Опишите основную идею парадигмы MapReduce

Критерии оценивания

- оценка «отлично (10)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений
- оценка «отлично (9)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений
- оценка «отлично (8)» выставляется студенту, показавшему всесторонние систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, и правильное обоснование принятых решений
- оценка «хорошо (7)» выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «хорошо (6)» выставляется студенту, если он знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «хорошо (5)» выставляется студенту, если он знает материал, и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «удовлетворительно (4)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;
- оценка «удовлетворительно (3)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет фрагментарно основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;
- оценка «неудовлетворительно (2)» выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач
- оценка «неудовлетворительно (1)» выставляется студенту, который не знает формулировок основных понятий дисциплины

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Дифференцированный зачет проводится по итогам текущей успеваемости и сдачи заданий и других видов работ, предусмотренных программой дисциплины и (или) путем организации специального опроса, проводимого в устной и (или) письменной форме.