Федеральное государственное образовательное бюджетное учреждение высшего образования
«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»
(Финансовый университет)

Департамент анализа данных, принятия решений и финансовых технологий

И.В.Синицын

ТЕХНОЛОГИИ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки
09.04.03 «Прикладная информатика»
направленности «Интеллектуальные информационные технологии в управлении финансами организации», «Интеллектуальные информационные технологии в экономике и финансах»

Москва 2017

Федеральное государственное образовательное бюджетное учреждение высшего образования
«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»

(Финансовый университет)

Департамент анализа данных, принятия решений и финансовых технологий

УТВЕРЖДАЮ
Ректор Финуниверситета
_____ М.А. Эскиндаров
27.06.2017 г.

И.В.Синицын

ТЕХНОЛОГИИ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки
09.04.03 «Прикладная информатика»
направленности «Интеллектуальные информационные технологии в управлении финансами организации», «Интеллектуальные информационные технологии в экономике и финансах»

Рекомендовано Ученым советом факультета «Прикладная математика и информационные технологии»
(*протокол № 45 от «20» июня 2017 г.*)

Одобрено Департаментом анализа данных, принятия решений и финансовых технологий
(протокол № 14 от «19» июня 2017 г.)

Москва 2017

УДК 330.4(073)
ББК
Г51

Рецензент: А.В. Чечкин, д. ф.-м.н., профессор департамента анализа данных, принятия решений и финансовых технологий

Г51 И.В.Синицын. «Технологии обработки больших данных». Рабочая программа дисциплины для студентов, обучающихся направлению подготовки 09.04.03 «Прикладная информатика», направленности «Интеллектуальные информационные технологии в управлении финансами организации», «Интеллектуальные информационные технологии в экономике и финансах». – М.: Финансовый университет, департамент анализа данных, принятия решений и финансовых технологий, 2017. – 31 с.

Дисциплина «Технологии обработки больших данных» является дисциплиной общенаучного модуля дисциплин направления подготовки 09.04.03 «Прикладная информатика», направленности «Интеллектуальные информационные технологии в управлении финансами организации», «Интеллектуальные информационные технологии в экономике и финансах».

Рабочая программа дисциплины содержит цели и задачи дисциплины, требования к результатам освоения дисциплины, содержание дисциплины, тематику практических занятий и технологии их проведения, формы самостоятельной работы, систему оценивания, учебно-методическое и информационное обеспечение дисциплины.

УДК 330.4(073)
ББК

Учебное издание

Иван Васильевич Синицын

Технологии обработки больших данных

Рабочая программа дисциплины

Компьютерный набор, верстка: И.В.Синицын

Формат 60x90/16. Гарнитура *Times New Roman*
Усл. п.л. 0,0. Изд. № 00.0 - 2017. Тираж - экз.

Заказ № _____

Отпечатано в Финансовом университете

Содержание

1. Наименование дисциплины

Технологии обработки больших данных

2. Перечень планируемых результатов обучения дисциплине

В совокупности с дисциплинами базовой части модуля математики и информатики дисциплина «Технологии обработки больших данных» обеспечивает формирование следующих компетенций:

способностью анализировать данные и оценивать требуемые знания для решения нестандартных задач с использованием математических методов и методов компьютерного моделирования (ПК-8)

знать методы и особенности алгоритмов обработки больших данных;

уметь применять перспективные методы и алгоритмы решения профессиональных задач обработки больших данных;

владеть навыками создания прикладных программ для работы с большими данными;

способностью применять современные методы и инструментальные средства прикладной информатики для автоматизации и информатизации решения прикладных задач различных классов и создания ИС (ПК-11)

знать принципы работы высоконагруженных систем, обеспечивающих процессы хранения и анализа больших данных;

уметь эксплуатировать и сопровождать техническое и программное обеспечение высоконагруженных систем;

владеть навыками установки и обслуживания систем, обеспечивающих процессы хранения и анализа больших данных;

способностью проектировать архитектуру и сервисы ИС предприятий и организаций в прикладной области (ПК-12)

знать основные элементы систем, реализующих процесс хранения и анализа больших данных и принципы их работы;

уметь выявлять нештатные ситуации и локализовывать неисправности при работе систем, обеспечивающих процессы хранения и анализа больших данных;

владеть навыками проектирования и внедрения систем, обеспечивающих процессы хранения и анализа больших данных, в прикладной области;

способностью проектировать информационные процессы и системы с использованием инновационных инструментальных средств, адаптировать современные ИКТ к задачам прикладных ИС (ПК-13)

знать состав систем, обеспечивающих процессы хранения и анализа больших данных, а также порядок инсталляции их компонентов;

уметь осуществлять настройку параметров систем, обеспечивающих процессы хранения и анализа больших данных;

владеть навыками проектирования информационных процессов с использованием систем, обеспечивающих процессы хранения и анализа больших данных.

3. Место дисциплины в структуре образовательной программы

Дисциплина «Технологии обработки больших данных» является дисциплиной общенаучного модуля направления подготовки 09.04.03 «Прикладная информатика», направленности «Интеллектуальные информационные технологии в управлении финансами организации», «Интеллектуальные информационные технологии в экономике и финансах»

Изучение дисциплины «Технологии обработки больших данных» базируется на знаниях, полученных в рамках изучения математических дисциплин, входящих в ОП бакалавра по направлению подготовки 09.04.03

«Прикладная информатика»: Математический анализ, Алгебра и геометрия, Теория вероятностей и математическая статистика, Дискретная математика.

Требования к входным знаниям, умениям и владениям студентов:

Знать:

основные определения и линейной алгебры, математического анализа, математической логики и теории вероятностей;

Уметь:

применять математические методы, изученные в указанных дисциплинах, для решения учебных задач.

Владеть:

навыками создания программ для решения вычислительных и оптимизационных задач.

4. Объём дисциплины и виды учебной работы

Вид учебной работы по дисциплине	**Всего (в з/е и часах)**	**Модуль 6** 2016 год набора	**Модуль 6** 2017 год набора
Общая трудоёмкость дисциплины	**144**	**144**	**144**
Аудиторные занятия	**48/40**	**48**	**40**
Лекции	**16/10**	**16**	**10**
Практические занятия и семинарские занятия, в т.ч.	**32/30**	**32**	**30**
занятия в интерактивной форме	**32/30**	**32**	**30**
Самостоятельная работа	**96/104**	**96**	**104**
Вид текущего контроля	**Контрольная работа**	**Контрольная работа**	**Контрольная работа**
Вид промежуточной аттестации	**Экзамен**	**Экзамен**	**Экзамен**

5. Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий

5.1. Содержание дисциплины

Тема 1. Основные принципы работы с большими данными

Горизонтальная масштабируемость, отказоустойчивость и локальность при работе с большими данными. Распределенная файловая система HDFS: архитектура, технология, примеры. MapReduce: основная парадигма, технология, примеры. NoSQL базы данных.

Тема 2. Анализ больших данных. Методы Data Mining. Платформа Hadoop

Основы Hadoop. Базовый набор компонентов Hadoop. Базовые модули фреймворка Apache Framework. Файловая система Hadoop Distributed File System (HDFS). MapReduce Framework и YARN. Окружение Hadoop. Базовые приложения Hadoop. Закономерности формирования BigTable. Особенности отображения временных меток в BigTable. HIVE – хранилище больших данных: архитектура, работа с данными.

Тема 3. Аналитика в «Больших данных»

Анализ данных с помощью PIG. Команды PIG. Логистический анализ с использованием Splunk. Работа с данными с помощью Spark DataFrames и Spark SQL.

Тема 4. Машинное обучение и «Большие данные»

Основы машинного обучения. «Осмысление данных» для машинного обучения. Классификация инструментов, техник и алгоритмов машинного обучения. Ассоциативные правила. Кластерный анализ.

5.2. Учебно-тематический план

Прием 2016 года

№ п/п	Наименование темы (раздела) дисциплины	Трудоемкость в часах						Формы текущего контроля успеваемости
		Всего	Аудиторная работа				Самостоятельная работа	
			Общая	Лекции	Практические и семинарские занятия	Занятия в интерактивных формах		
1.	Основные принципы работы с большими данными.	36	12	4	8	8	24	Самостоятельные работы. Участие в решении задач на практических занятиях. Собеседования по домашним заданиям.
2.	Анализ больших данных. Методы Data Mining. Платформа Hadoop.	36	12	4	8	8	24	
3.	Аналитика в «Больших данных».	36	12	4	8	8	24	
4.	Машинное обучение и «Большие данные».	36	12	4	8	8	24	
	Итого	**144**	**48**	**16**	**32**	**32/100%**	**96**	

Прием 2017 года

№ п/п	Наименование темы (раздела) дисциплины	Трудоемкость в часах						Формы текущего контроля успеваемости
		Всего	Аудиторная работа				Самостоятельная работа	
			Общая	Лекции	Практические и семинарские занятия	Занятия в интерактивных формах		
1.	Основные принципы работы с большими данными.	36	10	2	8	8	26	Самостоятельные работы. Участие в решении
2.	Анализ больших данных. Методы Data	36	10	2	8	8	26	

	Mining. Платформа Hadoop.							задач на практических занятиях. Собеседования по домашним заданиям.
3.	Аналитика в «Больших данных».	36	10	2	8	8	26	
4.	Машинное обучение и «Большие данные».	36	10	4	6	6	26	
	Итого	**144**	**40**	**10**	**30**	**30/100 %**	**104**	

5.3. Содержание практических занятий

1. Наименование темы (раздела) дисциплины

Основные принципы работы с большими данными.

Тема семинарского занятия

Развитие больших данных. Цикл Гартнера в развитии информационных технологий.

Содержание практического занятия

История возникновения термина «Большие данные». Источники больших данных. Развитие больших данных. Цикл Гартнера в развитии информационных технологий.

Интерактив – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений – до 50% от трудоемкости практического занятия.

Рекомендуемые источники: п.8, [1], [2].

2. Наименование темы (раздела) дисциплины

Основные принципы работы с большими данными

Тема семинарского занятия

Методы многомерного статистического анализа и анализа нечисловой информации

Содержание практического занятия

Простейшие статистические характеристики. Приведение к нормальной форме. Оцифровка нечисловых данных. Особенности анализа количественных и качественных признаков. Методы шкалирования.

Интерактив – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений – до 50% от трудоемкости практического занятия.

Рекомендуемые источники: п.8, [1], [2].

3. Наименование темы (раздела) дисциплины

Анализ больших данных. Методы Data Mining. Платформа Hadoop.

Тема семинарского занятия

Введение в Hadoop.

Содержание практического занятия

Общая концепция и архитектура Hadoop. Установка Hadoop.

Интерактив – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений – до 50% от трудоемкости практического занятия.

Рекомендуемые источники: п.8, [1], [2]

4. Наименование темы (раздела) дисциплины

Анализ больших данных. Методы Data Mining. Платформа Hadoop.

Тема семинарского занятия

Фреймворк Apache Framework.

Содержание практического занятия

Работа с базовыми модулями фреймворка Apache Framework.

Интерактив – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений – до 50% от трудоемкости практического занятия.

Рекомендуемые источники: п.8, [1], [2]

5. Наименование темы (раздела) дисциплины

Анализ больших данных. Методы Data Mining. Платформа Hadoop.

Тема семинарского занятия

Файловая система HDFS.

Содержание практического занятия

Распределенная файловая система Hadoop Distributed File System: архитектура, технология, примеры. Сравнительная характеристика производительности HDFS.

Интерактив – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений – до 50% от трудоемкости практического занятия.

Рекомендуемые источники: п.8, [1], [2]

6. Наименование темы (раздела) дисциплины

Анализ больших данных. Методы Data Mining. Платформа Hadoop.

Тема семинарского занятия

MapReduce Framework и YARN

Содержание практического занятия

Использование фреймворка YARN для управления ресурсами кластера и менеджмента задач.

Интерактив – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений – до 50% от трудоемкости практического занятия.

Рекомендуемые источники: п.8, [1]

7. Наименование темы (раздела) дисциплины

Анализ больших данных. Методы Data Mining. Платформа Hadoop.

Тема семинарского занятия

Окружение Hadoop.

Содержание практического занятия

Работа с окружением Hadoop: YARN, Tez, Spark, HRS.

Интерактив – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений – до 50% от трудоемкости практического занятия.

Рекомендуемые источники: п.8, [1]

8. Наименование темы (раздела) дисциплины

Анализ больших данных. Методы Data Mining. Платформа Hadoop.

Тема семинарского занятия

Базовые приложения Hadoop.

Содержание практического занятия

Работа с базовыми приложениями Hadoop: Pig, HIVE, HBASE.

Интерактив – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений – до 50% от трудоемкости практического занятия.

Рекомендуемые источники: п.8, [1], [2]

9. Наименование темы (раздела) дисциплины

Аналитика в «Больших данных».

Тема семинарского занятия

Анализ данных с помощью PIG. Команды PIG.

Содержание практического занятия

Технология MapReduce. Создание приложения для MapReduce. Изучение основ языка для написания потоков PIG.

Интерактив – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений – до 50% от трудоемкости практического занятия.

Рекомендуемые источники: п.8, [1], [2]

10. Наименование темы (раздела) дисциплины

Аналитика в «Больших данных».

Тема семинарского занятия

Логистический анализ с использованием Splunk.

Содержание практического занятия

Использование Splunk для логического анализа логов: создание, агрегирование, создание вычисляемых полей, построение графиков.

Интерактив – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений – до 50% от трудоемкости практического занятия.

Рекомендуемые источники: п.8, [1], [2]

11. Наименование темы (раздела) дисциплины

Аналитика в «Больших данных».

Тема семинарского занятия

Работа с данными с помощью Spark DataFrames.

Содержание практического занятия

Работа с распределенными коллекциями данных, организованных посредством именованных столбцов.

Интерактив – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений – до 50% от трудоемкости практического занятия.

Рекомендуемые источники: п.8, [1], [2]

12. Наименование темы (раздела) дисциплины

Аналитика в «Больших данных».

Тема семинарского занятия

Работа с данными с помощью Spark SQL.

Содержание практического занятия

Работа в Spark SQL: преобразование данных в более эффективные форматы, секционирование данных, выполнение оптимизации.

Интерактив – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений – до 50% от трудоемкости практического занятия.

Рекомендуемые источники: п.8, [1], [2]

13. Наименование темы (раздела) дисциплины

Машинное обучение и «Большие данные».

Тема семинарского занятия

Основы KNIME.

Содержание практического занятия

Создание Workflow. Создание и конфигурация узла и его диалогов. Работа с данными и построение графиков.

Интерактив – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений – до 50% от трудоемкости практического занятия.

Рекомендуемые источники: п.8, [1]

14. Наименование темы (раздела) дисциплины

Машинное обучение и «Большие данные».

Тема семинарского занятия

Кластеризация в KNIME

Содержание практического занятия

Отбор выборки объектов для кластеризации, Определение множества переменных, по которым будут оцениваться объекты в выборке, вычисление значений меры сходства.

Интерактив – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений – до 50% от трудоемкости практического занятия.

Рекомендуемые источники: п.8, [1]

15. Наименование темы (раздела) дисциплины

Машинное обучение и «Большие данные».

Тема семинарского занятия

Социальные сети и базы знаний как источники больших данных.

Содержание практического занятия

Решение задач по сбору и анализу данных, полученных из социальных сетей и баз знаний.

Интерактив – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений – до 50% от трудоемкости практического занятия.

Рекомендуемые источники: п.8, [1]

16. Наименование темы (раздела) дисциплины

Машинное обучение и «Большие данные».

Тема семинарского занятия

MapReduce и обработка крупномасштабных графов

Содержание практического занятия

Постановка задачи обработки крупномасштабных графов. Решение задачи нахождения кратчайшего пути в крупномасштабном графе.

Интерактив – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) и коллективное обсуждение решений – до 50% от трудоемкости практического занятия.

Рекомендуемые источники: п.8, [1], [2]

6. Учебно-методическое обеспечение для самостоятельной работы обучающихся по дисциплине

«Технологии обработки больших данных»

6.1. Формы внеаудиторной самостоятельной работы

При изучении дисциплины «Технологии обработки больших данных» обязательными являются следующие формы самостоятельной работы:

- разбор теоретического материала по пособиям и конспектам лекций;
- самостоятельное изучение указанных теоретических вопросов;
- решение задач по темам практических занятий;

- выполнение контрольной работы;
- подготовка к экзамену.

Наименование разделов, тем входящих в дисциплину	Формы внеаудиторной самостоятельной работы	Трудоёмкость в часах	Указание разделов и тем, отводимых на самостоятельное освоение обучающимися
Основные принципы работы с большими данными.	Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.	24/26	Решение задач.
Анализ больших данных. Методы Data Mining. Платформа Hadoop.	Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий.	24/26	Решение задач.
Аналитика в «Больших данных».	Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Работа с источниками и поиск информаций в Интернете. Выполнение домашних заданий.	24/26	Решение задач.
Машинное обучение и «Большие данные».	Работа с учебной литературой. Решение типовых задач. Разбор вопросов по теме занятия. Выполнение домашних заданий.	24/26	Решение задач.
	Итого	96/104	

6.2. Методическое обеспечение для аудиторной и внеаудиторной самостоятельной работы

Текущий контроль осуществляется в ходе учебного процесса и контроля самостоятельной работы студентов, по результатам выполнения контрольной работы. Основными формами текущего контроля знаний являются:

- обсуждение вопросов и задач, вынесенных в планах семинарских занятий;
- решение задач и их обсуждение;
- выполнение контрольных заданий и обсуждение результатов;

- контрольная работа.

Промежуточная аттестация проводится в форме экзамена.

Критерии балльной оценки различных форм текущего контроля успеваемости содержатся в соответствующих методических рекомендациях департамента анализа данных, принятия решений и финансовых технологий.

Примеры задач:

1. Написать программу MapReduce для подсчета средней длины слов во входных файлах.
2. Написать программу MapReduce для сортировки данных об объемах продаж, информация о которых расположена в текстовом отчете, и выдать время максимального объема продаж.
3. Написать программу для подсчета слов в нескольких больших файлах
4. Написать программу, объединяющую два больших отсортированных набора данных в один отсортированный набор

7. Фонд оценочных средств

7.1. Перечень компетенций с указанием этапов их формирования в процессе освоения образовательной программы

Перечень компетенций, формируемых в процессе освоения дисциплины содержится в разделе 2. «Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы».

7.2. Описание показателей и критериев оценивания компетенций, описание шкал оценивания

Способность анализировать данные и оценивать требуемые знания для решения нестандартных задач с использованием математических методов и методов компьютерного моделирования (ПК-8).

Оценка уровня сформированности компетенции

Показатели оценивания	Критерии оценивания компетенции	Шкала оценивания
Знать методы и особенности алгоритмов обработки больших данных. **Уметь** применять перспективные методы и алгоритмы решения профессиональных задач обработки больших данных. **Владеть навыками** создания прикладных программ для работы с большими данными.	**Знать** основные алгоритмы обработки больших данных. **Уметь** применять имеющиеся алгоритмя для решения задач обработки больших данных. **Владеть** навыками графического представления результатов обработки больших данных.	**Пороговый уровень**
	Знать особенности создания алгоритмов обработки больших данных. **Уметь** создавать алгоритмы решения задач анализа больших данных. Владеть навыками разработки прикладных программ для обработки больших данных.	**Продвинутый уровень**
	Знать основные языковые конструкции языков обработки и анализа больших данных. **Уметь** визуально представлять результаты обработки и анализа больших данных. **Владеть** навыками коллективной разработки проектов анализа данных.	**Высокий уровень**

Способность применять современные методы и инструментальные средства прикладной информатики для автоматизации и информатизации решения прикладных задач различных классов и создания ИС (ПК-11).

Оценка уровня сформированности компетенции

<table>
<tr><th>Показатели оценивания</th><th>Критерии оценивания компетенции</th><th>Шкала оценивания</th></tr>
<tr><td rowspan="3">Знать принципы работы высоконагруженных систем, обеспечивающих процессы хранения и анализа больших данных.
Уметь эксплуатировать и сопровождать техническое и программное обеспечение высоконагруженных систем
Владеть навыками установки и обслуживания систем, обеспечивающих процессы хранения и анализа больших данных.</td><td>Знать особенности работы систем обработки больших данных. Уметь инсталлировать высоконагруженные системы. Владеть навыками работы в системах обработки и анализа больших данных.</td><td>Пороговый уровень</td></tr>
<tr><td>Знать порядок конфигурирования высоконагруженных систем. Уметь инсталлировать и оптимизировать работу систем обработки и анализа больших данных. Владеть навыками резервного копирования данных в высоконагруженных системах.</td><td>Продвинутый уровень</td></tr>
<tr><td>Знать принципы оптимизации работы высоконагруженных систем. Уметь конфигурировать программную и аппаратную части высоконагруженных систем. Владеть восстановления данных в высоконагруженных системах.</td><td>Высокий уровень</td></tr>
</table>

Способность проектировать архитектуру и сервисы ИС предприятий и организаций в прикладной области (ПК-12)

Оценка уровня сформированности компетенции

Показатели оценивания	Критерии оценивания компетенции	Шкала оценивания
Знать основные элементы систем, реализующих процесс хранения и анализа больших данных и принципы их работы. **Уметь** выявлять нештатные ситуации и локализовывать неисправности при работе систем, обеспечивающих процессы хранения и анализа больших данных. **Владеть навыками** проектирования и внедрения систем, обеспечивающих процессы хранения и анализа больших данных, в прикладной области.	**Знать** состав и структуру систем, реализующих процесс хранения и анализа больших данных. **Уметь** пользоваться базами данных, оперирующими большими массивами данных. **Владеть** навыками установки и настройки систем, обеспечивающих процессы хранения и анализа больших данных.	**Пороговый уровень**
	Знать принципы работы средств разработки систем, обеспечивающих процессы хранения и анализа больших данных. **Уметь** локализовывать неисправности в системах, реализующих процесс хранения и анализа больших данных. **Владеть навыками** формирования требований к системам обеспечивающим процессы хранения и анализа больших данных.	**Продвинутый уровень**
	Знать принцыпы построения составных частей и фреймворков, а также надстроек систем, обеспечивающих процессы хранения и анализа больших данных. **Уметь** конфигурировать программные комплексы, осуществляющие взаимодействие с системами, реализующих процесс хранения и анализа больших данных.	**Высокий уровень**

	Владеть навыками объединения нескольких источников данных и систем, обеспечивающих процессы хранения и анализа больших данных.	

Способность проектировать информационные процессы и системы с использованием инновационных инструментальных средств, адаптировать современные ИКТ к задачам прикладных ИС (ПК-13).

Оценка уровня сформированности компетенции

Показатели оценивания	Критерии оценивания компетенции	Шкала оценивания
Знать состав систем, обеспечивающих процессы хранения и анализа больших данных, а также порядок инсталляции их компонентов. **Уметь** осуществлять настройку параметров систем, обеспечивающих процессы хранения и анализа больших данных; **Владеть навыками** проектирования информационных процессов с использованием систем, обеспечивающих процессы хранения и анализа больших данных	**Знать** состав систем, обеспечивающих процессы хранения и анализа больших данных. **Уметь** инсталлировать высоконагруженные системы. **Владеть** навыками работы в системах обработки и анализа больших данных.	**Пороговый уровень**
	Знать порядок инсталляции высоконагруженных систем. **Уметь** инсталлировать и оптимизировать работу систем обработки и анализа больших данных. **Владеть навыками** резервного копирования данных в высоконагруженных системах.	**Продвинутый уровень**
	Знать порядок инсталляции и конфигурирования, принципы оптимизации работы высоконагруженных систем. **Уметь** конфигурировать программную и аппаратную части	**Высокий уровень**

	высоконагруженных систем. **Владеть** восстановления данных в высоконагруженных системах.	

7.3. Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, владений

Теоретические вопросы для подготовки к экзамену

1. История возникновения термина «Большие данные».
2. Источники больших данных.
3. Развитие больших данных. Цикл Гартнера в развитии информационных технологий.
4. Основы Hadoop. Базовый набор компонентов Hadoop.
5. Базовые модули фреймворка Apache Framework.
6. Файловая система Hadoop Distributed File System (HDFS). HDFS и HDFS2.
7. MapReduce Framework и YARN.
8. Окружение Hadoop: YARN, Tez, Spark, HRS.
9. Базовые приложения Hadoop: Pig, HIVE, HBASE.
10. HBase – распределенная и масштабируемая база данных для работы с большими данными.
11. Сравнение HBase и HDFS.
12. Модель данных в HBase.
13. Закономерности формирования BigTable.
14. Особенности отображения временных меток в BigTable.
15. HIVE – хранилище больших данных: архитектура, работа с данными.
16. Анализ данных с помощью PIG. Команды PIG.
17. Логистический анализ с использованием Splunk.
18. Работа с данными с помощью Spark DataFrames и Spark SQL.

19. Основы машинного обучения. «Осмысление данных» для машинного обучения.
20. Основы KNIME.
21. Классификация инструментов, техник и алгоритмов машинного обучения.
22. Ассоциативные правила.
23. Кластерный анализ.
24. Кластеризация в KNIME.
25. Кластеризация в Spank.
26. Социальные сети и базы знаний как источники больших данных.
27. Подходы к обработке больших данных, структурированных в виде графов.
28. Open Source-решения для обработки больших объемов графовых данных.
29. Apache Giraph и GraphLab.
30. MapReduce и обработка крупномасштабных графов.

Типовые задачи (практические задания)

1. Имеется большой корпус документов. Задача – для каждого слова, хотя бы один раз встречающегося в корпусе, посчитать суммарное количество раз, которое оно встретилось в корпусе.

2. имеется csv-лог рекламной системы вида:

<user_id>,<country>,<city>,<campaign_id>,<creative_id>,<payment></p>

11111,RU,Moscow,2,4,0.3

22222,RU,Voronezh,2,3,0.2

13413,UA,Kiev,4,11,0.7

…

Необходимо рассчитать среднюю стоимость показа рекламы по городам России.

3. Написать программу нахождения всех записей с IP-адресом 123.123.123.123» в логах web-сервера.

4. Написать программу удаления любой колонки в csv-логах.

5. Написать программу вставки всех записи из лога в базу данных.

6. Имеется набор текстовых документов, необходимо посчитать, сколько слов встретилось от 1 до 1000 раз в наборе, сколько слов от 1001 до 2000, сколько от 2001 до 3000 и так далее.

7. Имеются логи двух web-серверов. Необходимо посчитать для каждого IP-адреса на какой из 2-х серверов он чаще заходил.

8. Имеются 2 лога. Первый лог содержит лог web-сервера, второй файл содержит соответствие URL-> Тематика. Для каждого IP-адреса необходимо рассчитать страницы какой категории с данного IP-адреса загружались чаще всего.

7.4. Методические материалы, определяющие процедуры оценивания знаний, умений и владений

Соответствующие приказы, распоряжения ректората о контроле уровня освоения дисциплин и сформированности компетенций студентов.

8. Перечень основной и дополнительной учебной литературы

а) основная:

1. Миркин, Б. Г. Введение в анализ данных [электронный ресурс]: учебник и практикум / Б. Г. Миркин. — М.: Юрайт, 2017. — 174 с. — ЭБС: Юрайт

2. Калинина, В.Н. Анализ данных: компьютерный практикум / В.Н. Калинина, В.И. Соловьев. — М.: КНОРУС, 2017. — 166 с.

б) дополнительная:

1. Майер-Шенбергер, В. Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим / В. Майер-Шенбергер. — М.: Манн, Иванов и Фербер, 2014. — 240 с. — Режим доступа: (http://www.mann-ivanov-ferber.ru/books/paperbook/big_data/)

2. Крылов, В. В. Большие данные и их приложения в электроэнергетике: от бизнес-аналитики до виртуальных электростанций / В. В.Крылов, С.В. Крылов. — Нобель Пресс, 2014.

3. Hilbert, M. Big Data for Development: From Information- to Knowledge Societies" / M. Hilbert.–SSRN Scholarly Paper No. ID 2205145). Rochester, NY: Social Science Research Network. – 2013

3. Hortonworks. 7 Key Drivers for the Big Data Market. – 2012

4. Big Data analytics: Future architectures, Skills and roadmaps for the CIO – 2011. – IDC/SAS

9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

1. http://hadoop.apache.org.

2. Сайт кафедры департамента анализа данных, принятия решений и финансовых технологий.

3. Электронная библиотека Финансового университета (ЭБ) http://elib.fa.ru/ (http://library.fa.ru/files/elibfa.pdf)

4. Электронно-библиотечная система Znanium http://www.znanium.com
5. Электронно-библиотечная система издательства «ЮРАЙТ» https://www.biblio-online.ru/
6. Научная электронная библиотека eLibrary.ru http://elibrary.ru

10.Методические указания для обучающихся по освоению дисциплины

Самостоятельная работа студентов проходит аудиторно и внеаудиторно. Организации самостоятельной работы служит календарно-тематический план изучения дисциплины. В этом плане указана тематика лекций, практических занятий, вопросы и задания для самостоятельного изучения.

При подготовке к лекции целесообразно предварительно познакомиться с ее содержанием по рекомендованным пособиям и выделить наиболее трудные вопросы. Во время лекций следует конспектировать содержание лекции. После занятий следует провести работу с конспектом: отредактировать записи, оформить конспект. При оформлении целесообразно выделять специальным образом названия тем и формулировки вопросов, основные определения, формулировки теорем и примеры. Сделанные записи нужно сверить с учебниками и учебными пособиями и в случае расхождений проконсультироваться с преподавателем.

Методические указания по проведению практических занятий

По структуре практические занятия следует разделить на учебные и контрольные.

● ***Учебные практические занятия*** структурно состоят из следующих компонент:

1) проверка наличия выполненного задания самостоятельной работы каждого студента;

выборочная проверка корректности выполнения домашнего задания;

3) разбор типичных ошибок, возникших в самостоятельной работе;

4) рассмотрение теоретических вопросов, связанных с текущим практическим занятием;

5) разбор методов выполнения практических заданий и решения задач;

6) корректировка заданий для самостоятельной работы студентов;

7) интерактивная форма – Практикум по решению задач по тематике занятия в малых группах (2-4 студента) – представляет собой решение списка задач, определенных преподавателем, в группе из небольшого количества студентов. В каждой группе есть «сильный» студент, который может выполнять функции консультанта и помощника преподавателю. Работа группы оценивается по количеству правильно решенных задач.

● Контрольные практические занятия структурно состоят из следующих компонент:

1) проверка наличия контрольной работы каждого студента;

2) разбор типичных ошибок, возникших при выполнении контрольной работы;

3) проведение аудиторной контрольной работы.

При подготовке к практическому занятию необходимо повторить или, если это требуется, изучить соответствующий теоретический материал. Во

время занятия нужно точно записывать формулировки решаемых задач, вопросы, указания преподавателя к решению и разбираемые решения. После занятий необходимо просмотреть записанные решения и восстановить в решениях имеющиеся пробелы. В случае затруднений отметить соответствующие задания и обратиться за консультацией к преподавателю. Практические занятия проходят, как правило, в интерактивной форме и преподаватель учитывает активность студентов, направленную на решение предложенных задач, и в поиске ответов на вопросы. Не следует бояться дать неверный ответ или допустить иную ошибку: исправление и анализ ошибок в режиме общения с преподавателем и сокурсниками в ходе практического занятия способствуют освоению учебного материала и предупреждают появление ошибок в дальнейшем.

На практических занятиях используется проблемно-деятельностный подход для решения практических задач. Сущность проблемно-деятельностного обучения заключается в том, что в процессе учебных занятий создаются специальные условия, в которых обучающийся, опираясь на приобретенные знания, мысленно и практически действует в целях поиска и обоснования наиболее оптимальных вариантов ее решения. Создается проблемная задача, студенты знакомятся с задачей, анализируют ее, выделяют лежащее в ее основе противоречие, создают и обосновывают модель своих возможных действий по разрешению проблемной ситуации, пробуют разрешить возникшую проблему на основе имеющихся у них знаний, выстраивают модель своих действий по ее решению.

Домашние задания следует выполнять регулярно при подготовке к практическим занятиям. В большинстве своем задания являются типовыми, и образцы их решения содержатся в рекомендованных пособиях, в материале лекций и практических занятий. Если то или иное задание вызвало затруднение необходимо обратиться к преподавателю на консультации или

ближайшем практическом занятии. Регулярность в выполнении домашних заданий — важный фактор освоения дисциплины. Даже небольшие отклонения от графика могут спровоцировать серьезное отставание и в дальнейшем — риск получения неудовлетворительных оценок в ходе текущей и промежуточной аттестации. Для выполнения домашних заданий следует завести отдельную тетрадь. Контроль за выполнением домашних заданий осуществляется в ходе практических занятий и выборочного собеседования.

Примеры домашних заданий:

1. На жестком диске размером 1TB, полностью заполненном текстовыми файлами, найти одинаковые копии файлов.
2. На жестком диске размером 1TB, полностью заполненном текстовыми файлами, найти файлы с заданными словосочетаниями (black list, фразы спама…).
3. Даны файлы, в которых находится информация об объемах продаж крупной сети продуктовых магазинов. Спрогнозировать рыночную ситуацию (объем продаж на следующий месяц).
4. Дан архив сообщений пользователей социальной сети за последний год. Необходимо показать все сообщение указанного пользователя.
5. Дан файл, в котором записаны все события, выполненные клиентами крупного сотового оператора (звонки, сообщения, посещения страниц в интернете). Найти все факты наступления событий, связанных с заданным абонентом.
6. Дан файл, в котором указаны списки товаров, покупаемых совместно друг с другом в крупном сетевом магазине электронной техники. Выдать рекомендации по совместной покупке с указанным товаром (группой товаров).

11.Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине

1. Информационно-образовательный портал Финансового университета. http://portal.ufrf.ru.

2. Программные компоненты Hadoop; фреймворк Apache Framework; файловая система Hadoop Distributed File System; окружение Hadoop: YARN, Tez, Spark, HRS; базовые приложения Hadoop: Pig, HIVE, HBASE; Spark DataFrames и Spark SQL.

12. Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине

Требуется доступ в компьютерный класс для выполнения заданий для самостоятельной работы.