

Федеральное государственное образовательное бюджетное учреждение
высшего образования
**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ
РОССИЙСКОЙ ФЕДЕРАЦИИ»**
(Финансовый университет)
Департамент анализа данных, принятия решений и финансовых технологий
Алтайский филиал
Кафедра «Учет и информационные технологии в бизнесе»

Разработчик: Макрушин С.В.
Составитель: Чечулин Н.В.

ОБРАБОТКА СТАТИЧЕСКИХ И ПОТОКОВЫХ БОЛЬШИХ ДАННЫХ

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки
09.04.03 «Прикладная информатика»
направленность программы магистратуры
«Интеллектуальные информационные технологии в экономике и финансах»

Экономика и управление информационными технологиями

Программа двух квалификаций

Барнаул 2024

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

**Департамент анализа данных, принятия решений и финансовых технологий
Алтайский филиал**

Кафедра «Учет и информационные технологии в бизнесе»

УТВЕРЖДАЮ

УТВЕРЖДАЮ:

**Директор Алтайского филиала
Финуниверситета
_____ В.А. Иванова**

«23» апреля 2024 г.

Разработчик: Макрушин С.В.

Составитель: Чечулин Н.В.

**ОБРАБОТКА СТАТИЧЕСКИХ И ПОТОКОВЫХ БОЛЬШИХ
ДАННЫХ**

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки

09.04.03 «Прикладная информатика»

направленность программы магистратуры

«Интеллектуальные информационные технологии в экономике и финансах»

Экономика и управление информационными технологиями

Программа двух квалификаций

Рекомендовано Ученым советом

Алтайского филиала Финуниверситета

(протокол № 10 от 23.04.2024 г.)

*Одобрено заседанием кафедры «Учет и информационные технологии в
бизнесе» (протокол № 9 от 23 апреля 2024 г.)*

Барнаул 2024

Рецензенты: **Феклин В.Г.**, кандидат физико-математических наук, доцент
Департамента анализа данных, принятия решений и финансовых технологий
Финансового университета

Макрушин С.В.

Обработка статических и потоковых больших данных. Рабочая программа для обучающихся по направлению подготовки 09.04.03 «Прикладная информатика» направленность программы магистратуры «Интеллектуальные информационные технологии в экономике и финансах» Экономика и управление информационными технологиями. Программа двух квалификаций - М.: Финансовый университет, Департамент анализа данных, принятия решений и финансовых технологий, 2024. - 22 с. Барнаул: Алтайский филиал Финуниверситета, Кафедра «Учет и информационные технологии в бизнесе», 2024.

Дисциплина «Обработка статических и потоковых больших данных» является дисциплиной общенаучного модуля направления подготовки «Прикладная информатика» направленности программы магистратуры «Интеллектуальные информационные технологии в экономике и финансах».

Рабочая программа учебной дисциплины содержит требования к результатам освоения дисциплины, программу, тематику практических и семинарских занятий и их проведения, формы самостоятельной работы, контрольные вопросы и систему оценивания, учебно-методическое обеспечение дисциплины.

Учебное издание

Макрушин С.В. Чечулин Н.В.

Обработка статических и потоковых больших данных

Рабочая программа учебной дисциплины
Компьютерный набор, верстка Макрушина С.В.

Формат 60х90/16. Гарнитура *Times New Roman*
Усл. п.л. 1,75. Изд. № 4.1-20167. Тираж 26 экз.

Заказ _____

Отпечатано в Финансовом университете

© Макрушин С.В.
© Финансовый университет, 2024

Содержание

1. Наименование дисциплины.....	5
2. Перечень планируемых результатов освоения образовательной программы с указанием индикаторов их достижения, соотнесенных с планируемыми результатами обучения по дисциплине.....	5
3. Место дисциплины в структуре образовательной программы.....	6
4. Объем дисциплины в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся....	6
5. Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий.....	7
5.1. Содержание дисциплины.....	7
5.2. Учебно - тематический план.....	9
5.3. Содержание семинаров, практических занятий.....	9
6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине.....	11
6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы.....	11
6.2. Перечень вопросов, заданий, тем для подготовки к текущему контролю.....	12
7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине.....	14
8. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины.....	19
9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины.....	20
10. Методические указания для обучающихся по освоению дисциплины.....	21
11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем.....	21
12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине.....	22

1. Наименование дисциплины

Обработка статических и потоковых больших данных

2. Перечень планируемых результатов освоения образовательной программы с указанием индикаторов их достижения, соотнесенных с планируемыми результатами обучения по дисциплине

Код компетенции	Наименование компетенции	Индикаторы достижения компетенции ¹	Результаты обучения (владения ² , умения и знания), соотнесенные с компетенциями/индикаторами достижения компетенции
ПКН-6	Способность анализировать и прогнозировать основные социально-экономические показатели, предлагать стратегические направления экономического развития на микро-, мезо- и макроуровнях	1. Применяет методический инструментальный системного анализа и моделирования экономических процессов для обоснования, внедрения инновационных разработок с целью получения конкурентных преимуществ и обеспечения опережающего роста на новых и развивающихся рынках. 2. Обосновывает перспективы изменений основных социально-экономических показателей и стратегические направления экономического развития на микро-, мезо- и макроуровнях.	

ПКН-3	Способность применять инновационные технологии, методы системного анализа и моделирования экономических процессов при постановке и решении экономических задач	<p>1. Применяет современные математические модели и информационные технологии для прогнозирования тенденций экономического развития, решения экономических задач на макро-, мезо- и микроуровнях, оценки последствий принимаемых управленческих решений.</p> <p>2. Ранжирует стратегические и тактические цели экономического развития на макро-, мезо- и микроуровнях; использует фактологические (статистические и экономико-математические) методы для проведения анализа</p>	•
-------	--	--	---

3. Место дисциплины в структуре образовательной программы

Дисциплина «Обработка статических и потоковых больших данных» является дисциплиной общенаучного модуля направления подготовки «Прикладная информатика», направленность программы магистратуры «Интеллектуальные информационные технологии в экономике и финансах». Экономика и управление информационными технологиями (программа двух квалификаций) направления 38.04.01 – Экономика.

Изучение дисциплины «Обработка статических и потоковых больших данных» основывается на сумме знаний, полученных студентами в процессе изучения базовых дисциплин профессионального цикла, а также полученных при обучении в бакалавриате. Для изучения данной дисциплины студент должен обладать базовыми знаниями в области информационных технологий и компьютерных программ,

навыками программирования на языке Python.

Студент должен обладать навыками работы с первоисточниками, обобщения и интерпретации полученной информации, четкого изложения своей точки зрения, работы в команде.

4. Объем дисциплины в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся

Общая трудоемкость дисциплины составляет 4 зачётные единицы.

Вид промежуточной аттестации - экзамен.

Вид текущего контроля - контрольная работа.

Заочная форма обучения

Вид учебной работы по дисциплине	Всего (в з/е и часах)	Модуль 6 (в часах)
Общая трудоёмкость дисциплины	4/144	144
<i>Контактная работа – Аудиторные занятия</i>	48	48
Лекции	16	16
Семинары, практические занятия	32	32
<i>Самостоятельная работа</i>	96	96
Вид текущего контроля	Контрольная работа	Контрольная работа
Вид промежуточной аттестации	Экзамен	Экзамен

Очная форма обучения, 2018 г.

Вид учебной работы по дисциплине	Всего (в з/е и часах)	Модуль 6 (в часах)
Общая трудоемкость дисциплины	4/144	144
<i>Контактная работа - Аудиторные занятия</i>	30	30
Лекции	10	10
Семинары, практические занятия	20	20
<i>Самостоятельная работа</i>	114	114
Вид текущего контроля	Контрольная работа	Контрольная работа
Вид промежуточной аттестации	Экзамен	Экзамен

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

5.1. Содержание дисциплины

Тема 1. Профилирование процессов обработки данных.

Большие данные - определение и причины возникновения задач обработки больших данных. В рамках темы рассматривается профилирование реализации алгоритмов на Python, принципы решения задачи оптимизации производительности алгоритма. Рассматривается проблема выбора типичных средств обработки данных, адекватных различным объемам данных. Принцип обработки данных на базе операций map / filter / reduce.

Тема 2. Библиотека NumPy.

В рамках темы рассматривается технологический стек Python для обработки и анализа данных, возможности Python как glue language, специфика библиотеки NumPy и ее роль в экосистеме Python. Организация массивов в NumPy: хранение данных, создание массивов, принципы реализации операций с едиными исходными данными. Универсальные функции и применение функций по осям в NumPy. Принцип распространения значений при выполнении операций в NumPy: общий алгоритм и примеры Маскирование и прихотливое индексирование в NumPy.

Тема 3. Библиотека Pandas.

В рамках темы рассматривается организация Pandas DataFrame и организация индексации для DataFrame и Series; применение универсальных функций и работа с пустыми значениями в Pandas; общая логика выполнения объединения данных из нескольких Pandas DataFrame. Операция GroupBy в Pandas DataFrame и реализация в ней подхода «разбиение, применение и объединение»

Тема 4. Параллельная обработка данных.

В рамках темы рассматривается специфика современного аппаратного обеспечения для обработки больших данных и проблема масштабируемости параллельных вычислений. Многопроцессорные архитектуры с общей и разделяемой памятью - специфика и сравнение.

Подходы к декомпозиции крупных вычислительных задач на подзадачи для параллельного исполнения. Модели параллельного программирования и их сочетаемость с архитектурами параллельных вычислительных систем. Специфика различия между потоками и процессами.

Проблема Global Interpreter Lock в Python и способы обхода ее ограничений. Модуль Python multiprocessing - назначение и основные возможности, API multiprocessing.Pool.

Тема 5. Библиотека Dask.

В рамках темы рассматривается библиотека для анализа больших объемов данных Python Dask, различные предлагаемые ей подходы к обработке данных. В частности, три ключевых структуры данных Dask: Dask.Array, Dask.DataFrame и Dask.Bag их специфика и принцип выбора структур данных при решении задач. Рассматривается граф зависимостей задач, как ключевая структура для организации параллельной обработки данных в Python Dask. Рассматривается принцип и примеры использования распараллеливание алгоритмов с помощью dask.delayed .

Рассматривается структура данных Dask.Array, специфика ее реализации и применения, процедура создания, поддерживаемые Dask.Array операции и ее отличия от NumPy ndarray. Рассматривается структура данных Dask.DataFrame, специфика ее реализации и применения, процедура создания, ограничения использования Dask.DataFrame. Рассматриваются операции мэппинга в Dask.DataFrame и операции Dask.DataFrame работающие со скользящим окном. Рассматривается структура данных Dask.Bag, специфика ее реализации и применения, процедура создания, поддерживаемые Dask.Bag операции. Организация вычислений с помощью Map / Filter / Reduce: общий принцип и специфика параллельной реализации обработки данных с помощью Dask.Bag.

5.2. Учебно - тематический план

№ п/п	Наименование тем (разделов) дисциплины	Трудоемкость в часах			Формы текущего
		Всего	Аудиторная работа	Самос-	

			Общая, в т.ч.:	Лекц ии	Семина- ры, практи- ческие занятия	Занятия в инте- рактивн ых формах	тояте- льная работа	контроля успевае- мости*
1.	Профилирование процессов обработки данных	20	6	2	4	1	14	УО, ППЗ
2.	Библиотека NumPy	20	6	2	4	1	14	УО, ППЗ
3.	Библиотека Pandas	20	6	2	4	1	14	УО, ППЗ
4.	Параллельная обработка данных	20	6	2	4	1	14	УО, ППЗ
5.	Библиотека Dask	64	6	2	4	7	58	УО, ППЗ
	В целом по дисциплине	144	30	10	20	11	114	Контрольн ая работа
	Итого в %					37 %		

*Сокращения в таблице: УО - устный опрос; ППЗ - проверка практических заданий

5.3. Содержание семинаров, практических занятий

Наименование тем (разделов) дисциплины	Перечень вопросов для обсуждения на семинарских, практических занятиях, рекомендуемые источники из разделов 8,9 (указывается раздел и порядковый номер источника)	Формы проведения занятий
Тема 1. Профилирование процессов обработки	<ul style="list-style-type: none"> Большие данные - причины возникновения задач обработки больших данных. Профилирование реализации алгоритмов на Python. Принципы решения задачи оптимизации 	Интерактивная форма, работа на компьютере

данных	<p>производительности алгоритма.</p> <ul style="list-style-type: none"> • Проблема выбора типичных средств обработки данных, адекватных различным объемам данных. <i>Рекомендуемые источники: основная - 8.1, 8.2; дополнительная - 8.1 - 8.3; 9.3, 9.4, 9.6</i> 	
Тема 2. Библиотека NumPy	<ul style="list-style-type: none"> • Технологический стек Python для обработки и анализа данных. • Специфика библиотеки NumPy и ее роль в экосистеме Python. • Организация массивов в NumPy: хранение данных, создание массивов, принципы реализации операций с едиными исходными данными. • Универсальные функции и применение функций по осям в NumPy. <i>Рекомендуемые источники: основная - 8.1, 8.2; дополнительная - 8.1 - 8.3; 9.9</i> 	Интерактивная форма, работа на компьютере
Тема 3. Библиотека Pandas	<ul style="list-style-type: none"> • Организация Pandas DataFrame. • Организация индексации для DataFrame и Series. • Применение универсальных функций для Pandas DataFrame. • Общая логика выполнения объединения данных из нескольких Pandas DataFrame. <i>Рекомендуемые источники: основная - 8.1, 8.2; дополнительная 8.1 - 8.3; 9.2, 9.10</i> 	Интерактивная форма, работа на компьютере
Тема 4. Параллельная обработка данных	<ul style="list-style-type: none"> • Специфика современного аппаратного обеспечения для обработки больших данных и проблема масштабируемости параллельных вычислений. • Многопроцессорные архитектуры с общей и разделяемой памятью - специфика и сравнение. • Специфика различия между потоками и процессами. • Проблема Global Interpreter Lock в Python и способы обхода ее ограничений. • Модуль Python multiprocessing - назначение и основные возможности, API multiprocessing.Pool. <i>Рекомендуемые источники: основная - 8.1, 8.2; дополнительная - 8.1 - 8.3; 9.3, 9.4, 9.6, 9.7, 9.8</i> 	Интерактивная форма, работа на компьютере
Тема 5. Библиотека Dask	<ul style="list-style-type: none"> • Ключевые структуры данных Dask: Dask.Array, Dask.DataFrame и Dask.Bag их специфика и принцип выбора структур данных при решении задач. • Граф зависимостей задач, как ключевая структура для организации параллельной обработки данных в Python Dask. • Структура данных Dask.Array, специфика ее реализации и применения, процедура создания. • Поддерживаемые Dask.Array операции и ее отличия от NumPy ndarray. • Структура данных Dask.DataFrame, специфика ее реализации и применения. • Процедура создания, ограничения использования Dask.DataFrame. • Структура данных Dask.Bag, специфика ее 	Интерактивная форма, работа на компьютере

	<p>реализации и применения, процедура создания, поддерживаемые Dask.Bag операции.</p> <ul style="list-style-type: none"> • Организация вычислений с помощью Map / Filter / Reduce: общий принцип и специфика параллельной реализации обработки данных с помощью Dask.Bag. <p><i>Рекомендуемые источники:</i> основная - 8.1, 8.2. дополнительная - 8.1-8.3; 9.2, 9.9, 9.10, 9.11</p>	
--	---	--

6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы

Наименование тем (разделов) дисциплины	Перечень вопросов, отводимых на самостоятельное освоение	Формы внеаудиторной самостоятельной работы*
Тема 1. Профилирование процессов обработки данных	<ul style="list-style-type: none"> • Принцип обработки данных на базе операций map / filter / reduce. 	РЛ, РЭИ, РАП
Тема 2. Библиотека NumPy	<ul style="list-style-type: none"> • Принцип распространения значений при выполнении операций в NumPy: общий алгоритм и примеры. • Маскирование и прихотливое индексирование в NumPy. 	РЛ, РЭИ, РАП
Тема 3. Библиотека Pandas	<ul style="list-style-type: none"> • Работа с пустыми значениями в Pandas. • Операция GroupBy в Pandas DataFrame и реализация в ней подхода «разбиение, применение и объединение». 	РЛ, РЭИ, РАП
Тема 4. Параллельная обработка данных	<ul style="list-style-type: none"> • Подходы к декомпозиции крупных вычислительных задач на подзадачи для параллельного исполнения. • Модели параллельного программирования и их сочетаемость с архитектурами параллельных вычислительных систем. 	РЛ, РЭИ, РАП
Тема 5. Библиотека Dask	<ul style="list-style-type: none"> • Принцип и примеры использования распараллеливание 	РЛ, РЭИ, РАП

	алгоритмов с помощью <code>dask.delayed</code> . • Операции <code>Dask.DataFrame</code> работающие со скользящим окном. • Операции мэппинга в <code>Dask.DataFrame</code>	
--	---	--

* Сокращения в таблице: **НТ** - Номера тем; **ФВСТ** - Формы внеаудиторной самостоятельной работы; **Т/Ч** - Трудоемкость в часах; **РЛ** - работа с литературой; **РЭИ** - работа с электронными источниками; **РАП** - разработка алгоритмов и программ.

6.2. Перечень вопросов, заданий, тем для подготовки к текущему контролю

Примерный перечень вопросов к контрольной работе

1. Проблема Global Interpreter Lock в Python и способы обхода ее ограничений
2. Технологический стек Python для обработки и анализа данных, Python как glue language, специфика библиотеки NumPy и ее роль в экосистеме Python
3. Организация массивов в NumPy: хранение данных, создание массивов, принципы реализации операций с едиными исходными данными
4. Универсальные функции и применение функций по осям в NumPy
5. Принцип распространения значений при выполнении операций в NumPy: общий алгоритм и примеры
6. Маскирование и прихотливое индексирование в NumPy
7. Организация Pandas DataFrame и организация индексации для DataFrame и Series
8. Применение универсальных функций и работа с пустыми значениями в Pandas
9. Объединение данных из нескольких Pandas DataFrame: общая логика и примеры
10. Операция GroupBy в Pandas DataFrame и реализация в ней подхода «разбиение, применение и объединение»
11. Модуль multiprocessing - назначение и основные возможности, API multiprocessing.Pool

Примеры заданий контрольных работ

Задача. Из названий компаний, заданных в файле `partners.txt` выделить первое слово, последнее слово (если оно отличается от первого) и остальные слова (для названий, состоящих из 3х и более слов). Словом считается любой выделенный разделителями набор символов, не относящихся к знакам препинания.

Найти по 2 самых длинных слова, относящихся к каждой из 3х групп. Сформировать из 6 найденных слов список. Распараллелить расчёт при помощи

dask.delayed.

Задача. Подсчитать, сколько раз во всех текстовых файлах, лежащих в all_k.zip встречаются названия продуктов, оформленные в виде названий в кавычках. Выполнить задание с использованием Dask, распараллелив процесс обработки данных.

Критерии балльной оценки текущего контроля успеваемости

Промежуточная аттестация проводится в форме экзамена. Оценка знаний студентов осуществляется в баллах с учетом: оценки за работу в семестре; оценки итоговых знаний в ходе экзамена.

На экзамене осуществляется комплексная проверка компетенций студентов путем компьютерного тестирования.

Оценивание студентов на экзамене осуществляется в соответствии с требованиями и критериями 100-балльной шкалы, установленными в Финансовом Университете. Учитываются результаты текущего контроля (максимальная оценка - 40 баллов) и знания, навыки и умения, непосредственно показанные студентами в ходе экзамена (максимальная оценка 60 баллов).

Распределение максимального числа баллов по видам работы:

№ п/п	Вид отчетности	Баллы (макс)
1	Аттестация	20
2	Работа в семестре	20
3	Экзамен (зачет)	60
	Итого:	100

7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине:

Перечень компетенций с указанием индикаторов их достижения в процессе освоения образовательной программы содержится в разделе 2. *«Перечень планируемых результатов освоения образовательной программы с указанием индикаторов их достижения, соотнесенных с планируемыми результатами обучения по дисциплине».*

**Типовые контрольные задания или иные материалы, необходимые для оценки
уровня сформированности компетенций, умений и знаний**

Код компетенции	Наименование компетенции	Примеры заданий для оценки сформированности компетенции
ПКН-6	Способность анализировать и прогнозировать основные социально-экономические показатели, предлагать стратегические направления экономического развития на микро-, мезо- и макроуровнях	1. Применяет методический инструментарий системного анализа и моделирования экономических процессов для обоснования, внедрения инновационных разработок с целью получения конкурентных преимуществ и обеспечения опережающего роста на новых и развивающихся рынках. 2. Обосновывает перспективы изменений основных социально-экономических показателей и стратегические направления экономического развития на микро-, мезо- и макроуровнях.
ПКН-3	Способность применять инновационные технологии, методы системного анализа и моделирования экономических процессов при постановке и решении экономических задач	1. Применяет современные математические модели и информационные технологии для прогнозирования тенденций экономического развития, решения экономических задач на макро-, мезо- и микроуровнях, оценки последствий принимаемых управленческих решений. 2. Ранжирует стратегические и тактические цели экономического развития на макро-, мезо- и микроуровнях; использует фактологические (статистические и экономико-математические) методы для проведения анализа и системных оценок

Примеры практико-ориентированных заданий

Задача обработки большого объема числовой информации финансовой организации, хранящейся в заданном файле формата hdf5

1. В массиве чисел, хранящихся в файле finance.hdf5, найти строку (вывести ее

индекс и содержащиеся значения), в которой более всего значений превышающих среднее значение по всему массиву. Для расчётов использовать `dask.array`

2. В массиве чисел, хранящихся в файле `finance.hdf5`, подсчитать количество строк, в которых более 600 значений больше среднего значения по всему массиву. Для расчётов использовать `dask.array`.
3. В массиве чисел, хранящихся в файле `finance.hdf5`, подсчитать количество значений, не отклоняющихся от среднего значения более чем на 3 стандартных отклонения. Для расчетов использовать `dask.array`

Задача обработки большого объема числовой информации финансовой организации, хранящейся в заданном файле формата csv

1. В `accounts*.csv` найти `id`, для которого в столбце `amount` встречается наибольшее количество значений кратных трем. Выполнить задание с использованием Dask, распараллелив процесс обработки данных
2. В `accounts*.csv` найти `id`, для которого сумма положительных значений в столбце `amount` наибольшая. Выполнить задание с использованием Dask, распараллелив процесс обработки данных
3. В `accounts*.csv` найти `id`, для которого в столбце `amount` встречается наибольшее количество значений между 1000 и 1500. Выполнить задание с использованием Dask, распараллелив процесс обработки данных

Примерный перечень контрольных вопросов к экзамену

1. Большие данные - определение и причины возникновения задач обработки больших данных
2. Специфика современного аппаратного обеспечения для обработки больших данных и проблема масштабируемости параллельных вычислений
3. Выбор типичных средств обработки данных, адекватных различным объемам данных; принцип обработки данных на базе операций `map / filter / reduce`
4. Многопроцессорные архитектуры с общей и разделяемой памятью - специфика и сравнение
5. Подходы к декомпозиции крупных вычислительных задач на подзадачи для

параллельного исполнения

6. Модели параллельного программирования и их сочетаемость с архитектурами параллельных вычислительных систем
7. Профилирование реализации алгоритмов на Python, принципы решения задачи оптимизации производительности алгоритма
8. Проблема Global Interpreter Lock в Python и способы обхода ее ограничений
9. Технологический стек Python для обработки и анализа данных, Python как glue language, специфика библиотеки NumPy и ее роль в экосистеме Python
10. Организация массивов в NumPy: хранение данных, создание массивов, принципы реализации операций с едиными исходными данными
11. Универсальные функции и применение функций по осям в NumPy
12. Принцип распространения значений при выполнении операций в NumPy: общий алгоритм и примеры
13. Маскирование и прихотливое индексирование в NumPy
14. Организация Pandas DataFrame и организация индексации для DataFrame и Series
15. Применение универсальных функций и работа с пустыми значениями в Pandas
16. Объединение данных из нескольких Pandas DataFrame: общая логика и примеры
17. Операция GroupBy в Pandas DataFrame и реализация в ней подхода «разбиение, применение и объединение»
18. Модуль multiprocessing - назначение и основные возможности, API multiprocessing.Pool
19. Различия между потоками и процессами, различие между различными планировщиками в Dask
20. Граф зависимостей задач - суть структуры данных, ее построение и использование в Dask
21. Три ключевых структуры данных Dask: их специфика и принцип выбора структуры данных при решении задач
22. Dask.Array - структура данных, специфика реализации и применения, процедура создания
23. Dask.Array - поддерживаемые операции и отличия от NumPy ndarray

- 24.Распараллеливание алгоритмов с помощью `dask.delayed` - принцип и примеры использования
- 25.Дополнительные параметры декоратора `dask.delayed` - назначение и примеры использования
- 26.Использование `dask.delayed` для объектов и операции над объектами `dask.delayed`, включая ограничения их использования
- 27.`Dask.DataFrame` - структура данных, специфика реализации и применения, процедура создания `Dask.DataFrame`
28. Ограничения использования `Dask.DataFrame` и операции мэппинга в `Dask.DataFrame`
29. Поддержка `Dask.DataFrame` операций работающих со скользящим окном
- 30.Совместное использование промежуточных результатов в `Dask`: принцип работы и примеры использования
- 31.`Dask.Bag` - структура данных, специфика реализации и применения, процедура создания `DaskBag`
- 32.Организация вычислений с помощью `Map / Filter / Reduce` : общий принцип и специфика параллельной реализации обработки данных в `Dask.Bag`
33. API `Dask.Bag` - функции мэппинга, фильтрации и преобразования
34. API `Dask.Bag` - функции группировки и свертки

Пример экзаменационного билета

Федеральное государственное образовательное бюджетное учреждение
высшего образования

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ
РОССИЙСКОЙ ФЕДЕРАЦИИ»**

Департамент анализа данных, принятия решений и
финансовых технологий

Обработка статических и потоковых больших данных. Рабочая программа для магистров,
обучающихся по направлению подготовки 09.04.03 «Прикладная информатика» профиль
«Интеллектуальные информационные технологии в экономике и финансах»

Дисциплина: «Обработка статических и потоковых больших данных»
Факультет Прикладной математики и информационных технологий Форма
обучения очная

Учебный 201_/201_ год _____ семестр
ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № ____

1. Маскирование и прихотливое индексирование в NumPy (30 баллов)
2. В массиве чисел, хранящихся в файле random.hdf5, найти строку (вывести ее индекс и содержащиеся значения), в которой более всего значений превышающих среднее значение по всему массиву. Для расчетов использовать dask.array (30 баллов)

Заместитель руководителя

Подготовил

Дата

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

Приказ от 23.03.2017 №0557/о «Об утверждении Положения о проведении текущего контроля успеваемости и промежуточной аттестации обучающихся по программам бакалавриата и магистратуры в Финансовом университете».

8. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины:

Основная литература:

1. Лесковец, Ю. Анализ больших наборов данных / Ю. Лесковец, А. Раджараман, Д. Д. Ульман ; перевод с английского А. А. Слинкина. — Москва : ДМК Пресс, 2016. — 498 с. — ЭБС ZNANIUM.com.— URL: <http://znanium.com/catalog/product/1027845> (дата обращения: 13.12.2019). — Текст : электронный.
2. Форман, Дж. Много цифр. Анализ больших данных при помощи Excel / Дж. Форман; перевод с английского А. Соколовой. — Москва: Альпина Пабlishер, 2016.

— 461 с.—ЭБС ZNANIUM.com.— URL: <http://znanium.com/catalog/product/551044>
(дата обращения: 13.12.2019) .— Текст : электронный.

Дополнительная литература:

1. Дадян, Э. Г. Данные: хранение и обработка : учебник / Э.Г. Дадян. — Москва : ИНФРА-М, 2019. — 205 с. — ЭБС ZNANIUM.com. — www.dx.doi.org/10.12737/textbook_5cf8c7f2b8cdb8.06963680. —URL: <http://znanium.com/catalog/product/989190> (дата обращения: 13.12.2019) — Текст : электронный
2. Дейт, К. Д. Введение в системы баз данных/ К. Д. Дейт.— 8-е изд. — Москва: Издательский дом «Вильямс», 2005. — 1328 с. — Текст : непосредственный.
3. Мартишин, С.А. Базы данных: Работа с распределенными базами данных и файловыми системами на примере MongoDB и HDFS с использованием Node.js, Express.js, Apache Spark и Scala: учебное пособие / С.А. Мартишин, В.Л. Симонов, М.В. Храпченко. — Москва: ИНФРА-М, 2019. — 235 с. — ЭБС ZNANIUM.com. - URL: <http://znanium.com/catalog/product/1018196> (дата обращения: 13.12.2019). — Текст : электронный.

9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины

1. Pyru 1.0.9 [Электронный ресурс]: сайт. - Режим доступа: <https://pypi.python.org/pypi/pyru>
2. Python Data Analysis Library [Электронный ресурс]: сайт. - Режим доступа: <http://pandas.pydata.org/>
3. Python Documentation [Электронный ресурс]: сайт. - Режим доступа: <http://python.org/doc/>
4. Python Standard Library [Электронный ресурс]: сайт. - Режим доступа: <https://docs.python.org/2/library/>
5. Scikit-learn Machine Learning in Python [Электронный ресурс]: сайт. - Режим доступа: <http://scikit-learn.org>
6. Официальный сайт продукта <https://www.python.org/>
7. Каталог курсов Интернет Университета Информационных Технологий

<http://www.intuit.ru/>

8. The Python Tutorial // <https://docs.python.org/3/tutorial/index.html>
9. NumPy User Guide // <http://docs.scipy.org/doc/numpy/user/index.html>
10. Pandas User Guide <http://pandas.pydata.org/pandas-docs/stable/>
11. Dask User Guide <https://docs.dask.org/en/latest/>
12. Электронно-библиотечная коллекция Springer Nature <http://www.library.fa.ru/resource.asp?id=608> SpringerLink
13. Электронная библиотека Финансового университета (ЭБ) <http://elib.fa.ru/>
(<http://library.fa.ru/files/elibfa.pdf>)
14. Электронно-библиотечная система BOOK.RU <http://www.book.ru>
15. Электронно-библиотечная система «Университетская библиотека ОНЛАЙН» <http://biblioclub.ru/>
16. Электронно-библиотечная система Znanium <http://www.znanium.com>
17. «Деловая онлайн библиотека» издательства «Альпина Паблишер»
<http://lib.alpinadigital.ru/en/library>
18. Электронно-библиотечная система издательства «Лань»
<https://ZZe.lanbook.com/>
19. Электронно-библиотечная система издательства «ЮРАЙТ»
<https://www.biblio-online.ru/>
20. Научная электронная библиотека eLibrary.ru <http://ZzeLibrary.ru>

10. Методические указания для обучающихся по освоению дисциплины.

Методические указания для обучающихся по освоению дисциплины (комплекс рекомендаций и разъяснений, позволяющий студенту оптимальным образом организовать процесс изучения учебного материала дисциплины) представлены в Учебно-организационном комплексе для дисциплин Департамента анализа данных, принятия решений и финансовых технологий, размещенном на странице Департамента анализа данных, принятия решений и финансовых технологий сайта Финансового университета.

При изложении лекции используется проблемный подход, что значительно расширяет предоставленный материал. На преподавательском диске находятся тексты лекций, материалы практических занятий, разбитых по темам. Там же приведены постановки задач, образцы программ решения типовых задач и справочные материалы.

Для получения доступа к облачному хранилищу студенты должны получить соответствующую ссылку от преподавателя.

При переходе к новой теме проводится тестирование, направленное на оценивание теоретических знаний. Помимо тестирования, может проводиться выборочный устный опрос студентов.

Практические навыки оцениваются путем разработки прикладных программ. Студенты должны самостоятельно и вовремя решать поставленные преподавателем задачи. Преподаватель должен отмечать и поощрять наиболее исполнительных студентов.

11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем

11.1 Комплект лицензионного программного обеспечения:

1. Windows, Microsoft Office,
2. Антивирус ESET Endpoint Security

11.2 Современные профессиональные базы данных и информационные справочные системы:

1. Справочная правовая система «Консультант Плюс».
2. Справочная правовая система «Гарант».
3. Информационно-образовательный портал Финансового университета.

11.3 Сертифицированные программные и аппаратные средства защиты информации - не предусмотрены

12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине.

Для проведения лекций и практических занятий необходима аудитория, оснащенная проектором и компьютерами с постоянным подключением к сети Интернет.