Федеральное государственное автономное образовательное учреждение высшего образования
"Национальный исследовательский университет
"Высшая школа экономики"

Факультет компьютерных наук
Департамент больших данных и информационного поиска

Рабочая программа дисциплины "Анализ Больших данных в SAS"

для образовательной программы «Финансовые технологии и анализ данных»
направления 01.04.02 уровень магистр

Разработчик(и) программы
Петровский М., к.ф.-м.н., доцент, Mikhail.Petrovsky@sas.com,
Локтева Л., Lyudmila.Lokteva@sas.com,
Александров М., Mikhail.Alexandrov@sas.com,
Ефимов А., Alexander.Efimov@sas.com.

[Введите Фамилию И.О., ученую степень, звание автора 2, электронный адрес]

Одобрена на заседании кафедры/департамента/подразделения [Введите название кафедры/департамента/школы]
«___»_____ 201_ г.
Руководитель департамента/Школы
В.В.Подольский _____ [подпись]

Утверждена Академическим советом образовательной программы
«___»_____ 201_ г., № протокола_____

Академический руководитель образовательной программы
А.А. Масютин _____ [подпись]

_____, 201_

1 Область применения и нормативные ссылки

Настоящая программа учебной дисциплины устанавливает требования к образовательным результатам и результатам обучения студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих дисциплину «Анализ больших данных в SAS», учебных ассистентов и студентов направления подготовки 01.04.02, обучающихся по образовательной программе «Финансовые технологии и анализ данных».

Программа учебной дисциплины разработана в соответствии с:

- Образовательным стандартом НИУ ВШЭ направления 01.04.02 «Прикладная математика и информатика», уровень подготовки: магистр;
- Образовательной программой «Финансовые технологии и анализ данных».

2 Цели освоения дисциплины

Целями освоения дисциплины «Анализ больших данных SAS являются:

- Изучение математических методов и подходов, используемых в программных системах обработки и анализа больших данных компании SAS - мирового лидера в области разработки и внедрения IT решений и услуг в бизнес-аналитике, в том числе в финансовой и банковской сфере.
- Развитие профессиональных навыков учащихся за счет получения практического опыта работы с IT решениями компании SAS в части обработки и анализа больших данных.

Достижение этих целей позволит подготовить высококвалифицированных специалистов, обладающих уникальной комбинацией компетенций и определенным практическим в передовых областях развития информационных технологий и математики, применяемых в настоящее время для решения задач анализа больших данных в ведущих финансовых организациях мира.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

Знать:

- Основные понятия и терминологию в области технологий обработки и анализа больших данных, понятие модели параллельной обработки данных MapReduce и ее практическую реализацию в Hadoop, основные API и типовые примеры программирования в Hadoop, стек Hadoop-технологий для распределенного хранения и обработки данных.
- Языки программирования для обработки данных SAS Base и SAS Data Step 2, а также методы организации взаимодействия аналитической платформы SAS с распределенными хранилищами информации на основе Hadoop-стека технологий.
- Семейство программных технологий SAS для обработки и анализа больших данных, включая программные продукты SAS для автоматической загрузки и предобработки больших данных, интерактивного исследования данных с использованием In-Memory подхода, построения прогнозных и описательных моделей, операционализации моделей, средств обработки разнородных сложно-структурированных данных, в том числе текстовых и сетевых (графовых).
- Математические методы и модели представления данных для решения задач машинного обучения, статистического и интеллектуального анализа данных, прогнозирования временных рядов, задач оптимизации, используемых в программных решениях компании SAS, в том числе для распределенной работы со сложно-структурированными данными большого объема.

Уметь:

- Разрабатывать программный код для эффективной обработки распределенных данных большого объема с использованием языков программирования SAS Base и SAS DS2 в сочетании с распределенным хранением данных в Hadoop кластере.
- Строить и применять на практике описательные и прогнозные модели интеллектуального анализа больших данных и машинного обучения с использованием технологий компании SAS, в том числе разнородных сложно-структурированных данных большого объема.
- Использовать программные средства визуализации и интерактивного исследования больших данных.

Владеть:

- Языками программирования SAS Base и SAS DS2 для обработки больших данных с помощью SAS ACCESS for Hadoop и PROC HADOOP.
- Программными средствами автоматической загрузки и обработки больших данных SAS Data Loader for Hadoop.
- Навыками работы с системой интерактивного исследования и визуализации больших данных SAS Visual Analytics и Visual Statistics.
- Навыками построения описательных и прогнозных аналитических моделей в системах SAS InMemory Statistics, SAS Enterprise Miner, SAS Text Miner, SAS Forecast Studio.
- Навыками решения задач оптимизации с использование пакета SAS OR (operations research).
- Навыками работы с сетевыми данными с использованием аналитических средств SAS SNA (social networks analysis).

Уровни формирования компетенций:

РБ — ресурсная база, в основном теоретические и предметные основы (знания, умения);

СД – способы деятельности, составляющие практическое ядро данной компетенции;

МЦ – мотивационно-ценностная составляющая, отражает степень осознания ценности компетенции человеком и готовность ее использовать

В результате освоения дисциплины студент осваивает компетенции:

Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
СК-1, СК-4, СК-8	МЦ	дает определение, воспроизводит, распознает, использует	Лекции и практические занятия	Практические задания, экзамен
ПК11, ПК14,	РБ	владеет, применяет	Лекции и практические занятия	Практические задания, экзамен
ПК17, ПК18, ПК20	СД	владеет, применяет, обосновывает, интерпретирует, оценивает	Лекции и практические занятия	Практические задания, экзамен

4 Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к циклу элективных дисциплин образовательной программы «Финансовые технологии и анализ данных».

Изучение данной дисциплины базируется на следующих дисциплинах:

- Теория баз данных
- Эконометрика
- Машинное обучение

5 Тематический план учебной дисциплины

[Тематический план отражает содержание дисциплины (перечень разделов), структурированное по видам учебных занятий с указанием их объемов в соответствии с ОУП]
[Таблица для дисциплин, закрепленных за одной кафедрой/подразделением[1]]

№	Название раздела	Всего часов	Аудиторные часы				Самостоятельная работа
			Лекции	Семинары	Практические занятия	Другие виды работы[2]	
	Раздел 1. Технологии хранения и обработки Больших данных		**8**	**0**	**12**		**22**
1	Введение в Большие данные		1	0	0		0
2	Модель вычислений MapReduce		2	0	4		8
3	Реализации алгоритмов на MapReduce.		2	0	4		8
4	Основы Hive.		2	0	4		6
	Раздел 2. Программирование обработки и загрузки Больших данных в SAS		**10**		**20**		**30**
5	Основы языка SAS Base		1	0	4		6
6	Макропроцессор SAS. SAS SQL.		1	0	4		6
7	Обзор SAS Data Loader for Hadoop		1	0	4		6
8	Взаимодействие SAS и Hadoop.		1	0	4		6
9	Основы программирования на SAS Data Step 2		1	0	4		6
	Раздел 3. Аналитика в больших данных		**10**	**0**	**10**		**30**
10	Обзор задач и методов машинного обучения и интеллектуального анализа больших данных		2	0	2		6
11	Интерактивное исследование данных с помощью SAS VA/VS		2	0	2		6
12	Построение моделей с помощью IMSTAT в		2	0	2		6

[1] Также программа может быть разработана департаментом, школой, институтом или другим подразделением НИУ ВШЭ, реализующими учебную дисциплину

[2] Указать другие виды аудиторной работы студентов, если они применяются при изучении данной дисциплины.

	SAS LASR сервере						
13	Интеллектуальный анализ данных с SAS Enterprise Miner		2	0	2		6
14	Операционализация моделей		2	0	2		6
	Раздел 4. Аналитическая обработка сложно-структурированных больших данных		**8**	**0**	**8**		**24**
15	Обработка текстов в SAS Text Miner		2	0	3		6
16	Анализ временных рядов в SAS Forecast Studio.		2	0	3		6
17	Анализ взаимосвязей с помощью SAS SNA.		2	0	3		6
18	Решение оптимизационных задач в SAS/OR		2	0	3		6
	Общий итог		**30**	**0**	**54**		**106**

6 Формы контроля знаний студентов

Тип контроля	Форма контроля	1 год		Параметры **
		3	4	
Текущий	Лабораторная работа (практическое задание)	*	*	Демонстрация программной реализации или программного проекта, отчет
	Домашнее задание (практическое задание)	*	*	Программная реализация или программный проект, отчет
Итоговый	Экзамен		*	письменный экзамен 90 мин.

7 Критерии оценки знаний, навыков

Оценки по всем формам текущего контроля выставляются по 10-ти балльной шкале. Промежуточный контроль в форме домашнего задания и лабораторной работы подразумевает самостоятельное выполнение студентом поставленной практической задачи с использованием указанного преподавателем программного средства в указанные сроки.

Домашние задания выполняются с использованием удаленного доступа или локальной установки необходимого программного обеспечения SAS и последующей отсылкой преподавателю решения в виде программного или программного проекта в соответствующей среде, а также краткого отчета о выполнении данного кода, включающего как сгенерированные автоматически результаты выполнения программы, так и по необходимости текстовые комментарии студента.

Лабораторные работы выполняются в компьютерном классе с использованием удаленного доступа к необходимому программному обеспечению SAS. Работоспособность и правильность программной реализации поставленной преподавателем задачи демонстрируется непосредственно на занятии, также после занятия может по требованию преподавателя формироваться отчет о решенной задаче и пересылаться преподавателю.

Итоговый контроль представляет собой письменный экзамен, включающий малтичойс вопросы и задачи, для решения которых требуется написать соответствующий программный код или фрагмент кода.

8 Содержание дисциплины

Количество часов аудиторной работы – по темам согласно пункту 5.

Тема	Краткое содержание темы	Литература
Раздел 1. Технологии хранения и обработки Больших данных		[1],[2],[7]
Тема 1.1. Введение в Большие данные	Терминология, история появления. Технические сложности работы с большими данными. Распределенная файловая система HDFS. Базовая идея модели MapReduce, пример использования MapReduce.	
Тема 1.2. Модель вычислений MapReduce	Модель вычислений MapReduce. Реализация MapReduce в Hadoop. Основы Java API. Планирование вычислений. YARN.	

Тема 1.3. Реализации алгоритмов на MapReduce. Основы Hive.	Примеры реализации алгоритмов на MapReduce (включая умножение матриц, операции реляционной алгебры). Колоночные форматы хранения (на примере ORC). Основы Hive.	
Раздел 2. Программирование обработки и загрузки Больших данных в SAS		[1],[4-6]
Тема 2.1. Основы языка SAS Base	Изучение основ программирования на языке SAS: структуры языка, запуска и отладки программы, простейших аналитических процедур и приёмов для трансформации данных.	
Тема 2.2. Макропроцессор SAS. SAS SQL.	Изучение основ программирования на языке SAS Macro: использование макропрограмм и макропеременных для написания программ SAS со сложной структурой и логикой.	
Тема 2.3. Обзор SAS Data Loader for Hadoop: автоматизированная загрузка и обработка больших данных.	Hadoop как технология хранения и обработки больших данных. Способы загрузки данных в Hadoop. Базовые операции с данными. Профилирование, дедупликация. Выполнение процессов Data Quality внутри Hadoop. Интеграция с SAS In-Memory Analytics Server.	
Тема 2.4. Взаимодействие SAS и Hadoop.	Обзор взаимодействия SAS и HADOOP. Интерфейс SAS ACCESS в HADOOP (HIVE). Выполнение выражений Pig, HDFS, Map Reduce через PROC HADOOP. Интерфейс SAS ACCESS в HADOOP (Impala)	
Тема 2.5. Основы программирования на SAS Data Step 2 (DS2)	Изучаются основы программирования на языке DS2, который является специализированным языком программирования четвёртого поколения и используется в SAS для работы с данными.	
Раздел 3. Аналитика в больших данных		[1-3],[7],[9]
Тема 3.1. Обзор задач и методов машинного обучения и интеллектуального анализа больших данных	Аналитические методы, применимые к большим объёмам данных. Многомерные связи, ассоциации, корреляции. Непрерывность поступающих обновлений как характерная черта анализа больших данных. Примеры обработки неструктурированных данных. Понятия неоднозначности и недостоверности данных.	
Тема 3.2. Интерактивное исследование данных с помощью SAS VA/VS	Примеры визуализаций базового статистического анализа, доступных в SAS VA/VS. Агрегирование и частотный анализ. Диаграммы переходов в системах с процессами. Кластеризация и факторный анализ. Базовая диагностика моделей. Пример проекта - добавление данных, дизайн визуализаций.	
Тема 3.3. Построение моделей с помощью IMSTAT в SAS LASR сервере	Обзор аналитического сервера in-memory аналитики SAS LASR. Загрузка данных на LASR сервер. Исследование данных. Подготовка данных. Обзор алгоритмов машинного обучения в IMSTAT.	

Тема 3.4. SAS Enterprise Miner	Базовый функционал SAS EM - определение простого проекта, загрузка и изменение данных. Регрессионные модели и деревья решений. Поиск стандартных путей по истории процессов, частотный анализ и выявление отклонений. Диагностика и оценка качества моделей. Примеры: кредитные модели и оценка рисков.	
Тема 3.5. Операционализация моделей	Понятие операционализации моделей. Постановка моделей на регламентное переобучение и применение. Мониторинг качества моделей. Создание шаблонов построения моделей. Автоматическое построение моделей по сегментам данных. Экспорт и применение моделей в СУБД. Применение моделей в инструментах аналитики в реальном времени (SAS Decision Manager, SAS Event Stream Processing).	
Раздел 4. Аналитическая обработка сложноструктурированных больших данных		[1],[8-9]
Тема 4.1. Обработка текстов в SAS Text Miner	Применения технологий текстовой аналитики: обогащение информации по клиентам компаний, мониторинг потоков сообщений. Принципы статистического анализа текста (Text Mining): лингвистическая предобработка текста, статистическая фильтрация текста, автоматическое выявление тематик, кластеризация текстов.	
Тема 4.2. Анализ временных рядов (SAS Forecast Studio).	Использование интерактивного интерфейса SAS Forecast Studio для создания и использования прогнозных моделей для временных рядов, в том числе для автоматического создания и построения прогнозных моделей для данных большого объёма, содержащих временные ряды.	
Тема 4.3. Анализ взаимосвязей с помощью SAS SNA.	Терминология. Процедура PROC OPTGRAPH: разбиение графа, расчет метрик центральности и другие алгоритмы теории графов, анализа сетей и оптимизации.	
Тема 4.4. Решение оптимизационных задач в SAS/OR	Формулировка и решение задач линейного, нелинейного и целочисленного программирования с помощью процедуры OPTMODEL.	

9 Образовательные технологии

Основными формами преподавания в рамках данного курса являются лекции преподавателя, практические задания в форме домашних заданий и лабораторных работ, а также возможность использовать по некоторым темам дополнительные интерактивные электронные курсы компании SAS на английском языке.

10 Оценочные средства для текущего контроля и аттестации студента

10.1 Примеры заданий промежуточной аттестации

Задача (Тема 2.5). Дан набор данных заданной структуры и программа SAS Data step, производящая определенную обработку и вычисления с использованием данного набора. Перепишите эту программу на SAS DS2 с использованием параллельных нитей и созданием пользовательского пакета, чтобы результат обработки сохранился тем же, но код мог выполняться в параллельной среде.

Задача (Тема 3.3). Дан набор заданной структуры, постройте модель прогнозирования отклика с использованием процедуры impstat с алгоритмом random forest с заданным числом деревьев. Примените полученную модель к тестовому набору данных той же структуры, визуализируйте полученный график Lift. Постройте на том же наборе модель с использование высокопроизводительной версии метода GLM. Примените к тестовому набору. Сравните результаты GLM и Random Forest по AUC.

Задача (Тема 4.1). Дан текстовый корпус документов, лежащих в указанной директории. Создайте в SAS Tex Miner проект, который: выберет файлы с расширением pdf; осуществит парсинг набора с определением частей речи и сохранением в признаковом пространстве только существительных и глаголов; осуществит фильтрацию документов и признаков с использованием заданной схемой определения весов лексем (например, на основе tf-idf); выделит заданное количество ключевых тематик по методу SVD. В ответе укажите топ 5 ключевых слов во второй выявленной тематике. Какой документ имеет наибольший вес в это тематике?

11 Порядок формирования оценок по дисциплине

Формирование оценок по учебной дисциплине производится в соответствии с положением об организации промежуточной аттестации и текущего контроля успеваемости студентов, утвержденного Ученым советом НИУ ВШЭ от 27.06.2014, протокол №5.

В соответствии с учебным планом, формами текущего контроля являются лабораторная работа и домашнее задание. Каждая из форм текущего контроля оценивается по 10-балльной шкале. Общая оценка за текущий контроль (по 10-балльной шкале) рассчитывается по формуле:

$$O_{текущий} = (n_{дз} \cdot O_{дз} + n_{лр} \cdot O_{лр})/(n_{дз} + n_{лр})$$

Итоговая форма контроля – письменный экзамен, который также оценивается по 10-бальной шкале, пропорционально числу правильно решенных студентом задач и правильно отверченных вопросов. В итоговом экзамене студент получает индивидуальный набор задач и вопросов только по тем темам, по которым он получил за домашние и практические работы оценку 7 и ниже. Остальные темы считаются в экзамене сданными на 10 баллов. Результирующая оценка формируется по формуле:

$$O_{результ} = 0.7 * O_{текущий} + 0.3 * \cdot O_{экз}$$

Везде способ округления накопленной оценки текущего контроля – арифметический, в пользу студента.

12 Учебно-методическое и информационное обеспечение дисциплины

12.1 Базовый учебник

В качестве базовых учебных материалов по каждой лекции будет предоставляться учебная литература и наборы данных Учебного центра компании SAS персонально каждому слушателю. Электронная версия базового учебника не доступна.

12.2 Основная литература

1. Базовый учебник (пункт 12.1).
2. Чак Лэм Hadoop в действии - ДМК Пресс, 2012.
3. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. – Springer, 2001.
4. Lora D. Delwiche, Susan J. Slaughter The Little SAS® Book: A Primer, Fifth Edition - SAS Institute, 2012
5. Ron Cody An Introduction to SAS® University Edition - SAS Institute, 2015
6. Geoff Der, Brian S. Everitt Essential Statistics Using SAS® University Edition - SAS Institute, 2015
7. Jared Dean Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners - John Wiley Sons Inc, 2014
8. Tim Rey, Arthur Kordon, Chip Wells Applied Data Mining for Forecasting Using SAS® - SAS Institute, 2012
9. Kattamuri S. Sarma Predictive Modeling with SAS® Enterprise Miner™: Practical Solutions for Business Applications, Third Edition - SAS Institute, 2017
10. Anders Milhoj Practical Time Series Analysis Using SAS® - SAS Institute, 2013
11. Goutam Chakraborty, Murali Pagolu, Satish Garla Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS® - SAS Institute, 2013

12.3 Программные средства

Для успешного освоения дисциплины, студент использует следующие программные средства:

- Microsoft Azura HDInsight;
- SAS University Edition;
- SAS OnDemand for Academics (Enterprise Miner, Text Miner, Forecast Server, Enterprise Guide + SAS OR + SAS ETS);
- SAS Access for Hadoop;
- SAS SNA;
- SAS Data Loader for Hadoop, SAS Event Stream Processing;
- SAS LASR Server (Visual Analytics, Visual Statistics, InMemory Statistics);
- SAS Decision Manager.

12.4 Дистанционная поддержка дисциплины

- SAS Programming I: Essentials https://support.sas.com/edu/schedules.html?ctry=us&id=277;
- SAS Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression https://support.sas.com/edu/schedules.html?ctry=us&id=1979.

13 Материально-техническое обеспечение дисциплины

Используются персональный компьютер (ноутбук) и проектор для проведения лекций и практических занятий, техническое оснащение компьютерных классов для выполнения лабораторных работ, удаленный доступ к дата центрам компании SAS для доступа к программному обеспечению, недоступному для локальной установки студентами или в компьютерных классах.