

Документ: Федеральное государственное автономное образовательное учреждение высшего образования
Информация о владельце: "Самарский государственный экономический университет"
ФИО: Кандрашина Елена Александровна
Должность: И.о. ректора ФГАОУ ВО «Самарский государственный экономический университет»
Дата подписания: 13.07.2026 14:11:32
Уникальный программный ключ:
2db64eb9605ce27edd3b8e8fdd32c70e0674ddd2

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ) «ТЕХНОЛОГИИ БОЛЬШИХ ДАННЫХ»

Уровень высшего образования: бакалавриат

Направление подготовки: 09.03.03 Прикладная информатика

Направленность (профиль) подготовки: Интеллектуальные цифровые системы и сервисы в управлении

Квалификация (степень) выпускника: бакалавр

Форма обучения: очная

Год набора (приема на обучение): 2026

Срок получения образования: 4 года

Объем:
в зачетных единицах: 4 з.е.
в академических часах: 144 ак.ч.

г. Самара, 2026

Разработчики:

Арапова Ю. В.

Рабочая программа дисциплины (модуля) составлена в соответствии с требованиями ФГОС ВО по направлению подготовки 09.03.03 Прикладная информатика, утвержденного приказом Минобрнауки от 19.09.2017 № 922, с учетом трудовых функций профессиональных стандартов: "Менеджер по информационным технологиям", утвержден приказом Минтруда России от 30.08.2021 № 588н; "Специалист по информационным системам", утвержден приказом Минтруда России от 13.07.2023 № 586н; "Руководитель проектов в области информационных технологий", утвержден приказом Минтруда России от 27.04.2023 № 369н; "Системный аналитик", утвержден приказом Минтруда России от 27.04.2023 № 367н.

Согласование и утверждение

№	Подразделение или коллегиальный орган	Ответственное лицо	ФИО	Виза	Дата, протокол (при наличии)
1	Кафедра прикладной информатики	Заведующий кафедрой, руководитель подразделения, реализующего ОП	Ермолина Л. В.	Рассмотрено	21.05.2026, № 9

1. Цель и задачи освоения дисциплины (модуля)

Цель освоения дисциплины - формирование у обучающихся теоретических знаний и практических навыков в области проектирования, разработки и сопровождения информационных систем, основанных на технологиях больших данных, для автоматизации задач организационного управления и бизнес-процессов, а также способности выполнять и управлять работами по созданию, модификации и индустриальной эксплуатации таких систем.

Задачи изучения дисциплины:

- Освоение архитектурных концепций и методов быстрого прототипирования Big Data ИС – изучение архитектур Data Lake, Data Lakehouse и гибридных подходов, компонентов распределённого хранения и обработки (HDFS, Apache Spark, MapReduce), инструментов приёма потоковых и пакетных данных (Apache Kafka, Apache NiFi), а также принципов и инструментов прототипирования на базе Python-экосистемы (PySpark, pandas) и облачных сервисов для решения аналитических задач и поддержки принятия решений в организационном управлении;
- Изучение методологий проектирования и разработки хранилищ больших данных – формирование умений выполнять концептуальное, логическое и физическое моделирование данных для распределённых систем, включая многомерные схемы, модель Data Vault и денормализацию, работать с реляционными аналитическими СУБД и NoSQL-хранилищами (колоночными, документными, графовыми, «ключ-значение»), проектировать слои Data Lake, строить ETL/ELT-пайплайны и витрины данных, а также управлять метаданными и версионированием данных (Delta Lake, Apache Atlas);
- Приобретение навыков обеспечения и контроля качества данных и управления процессами сопровождения Big Data ИС – выработка практических навыков профилирования, валидации и автоматизированного тестирования качества данных с применением инструментов экосистемы больших данных (Great Expectations, dbt, PySpark), внедрения практик CI/CD и DataOps для конвейеров обработки данных, мониторинга дрейфа данных и схем, а также разработки регламентов, чек-листов и документации для аудита качества в соответствии с корпоративными стандартами при выполнении и управлении работами по созданию, модификации и индустриальному сопровождению ИС.

2. Планируемые результаты обучения по дисциплине (модулю), соотнесенные с планируемыми результатами освоения образовательной программы

Компетенции, индикаторы и результаты обучения

ПК-1 Способен выполнять работы и управлять работами по созданию (модификации) и сопровождению ИС, автоматизирующих задачи организационного управления и бизнес-процессы

ПК-1.1 Разрабатывает архитектуру и прототип ИС в рамках выполнения работ и управления работами по созданию (модификации) и сопровождению ИС, автоматизирующих задачи организационного управления и бизнес-процессы

Знать:

ПК-1.1/Зн1 Типовые архитектуры информационных систем, ориентированных на обработку больших данных в задачах организационного управления, включая концепции «Data Lake», «Data Lakehouse» и гибридные подходы, объединяющие слои приёма потоковых и пакетных данных (Apache Kafka, Apache NiFi), распределённого хранения (HDFS, объектные хранилища S3), массово-параллельной обработки (Apache Spark, MapReduce), аналитических и BI-инструментов; жизненный цикл создания и модификации Big Data ИС; принципы и инструменты быстрого прототипирования систем больших данных с использованием Python-экосистемы (PySpark, pandas, библиотеки визуализации) и облачных сервисов, а также способы интеграции источников данных большого объёма, скорости и разнообразия в контур автоматизации бизнес-процессов для решения задач аналитики, отчётности и поддержки принятия решений;

Уметь:

ПК-1.1/Ум1 Разрабатывать архитектуру Big Data ИС, декомпозируя её на функциональные подсистемы (сбора, потоковой обработки, распределённого хранения, аналитических витрин, визуализации), создавать прототипы таких систем на Python, проводить отладку и оценивать масштабируемость прототипа на тестовых наборах данных, а также управлять работами по итеративной доработке архитектуры в соответствии с требованиями к модификации и сопровождению ИС, автоматизирующей задачи организационного управления;

Владеть:

ПК-1.1/Нв1 Навыками проектирования архитектур и быстрого прототипирования информационных систем больших данных с использованием стека технологий Spark и Python, методами визуального моделирования компонентов архитектуры (диаграммы потоков данных, схемы «Data Lake») с учётом внедрения механизмов распределённой обработки и хранения больших объёмов бизнес-информации, а также навыками координации работ по сборке, нагрузочному тестированию и оценке эффективности прототипов в рамках задач создания, модификации и сопровождения ИС организационного управления.

ПК-1.2 Проектирует, создает дизайн и разрабатывает базы данных ИС в рамках выполнения работ и управления работами по созданию (модификации) и сопровождению ИС

Знать:

ПК-1.2/Зн1 Принципы проектирования и дизайна хранилищ данных для Big Data ИС, включая реляционные аналитические СУБД, NoSQL-системы классов колоночных, документных, графовых и хранилищ типа «ключ-значение» в контексте работы с большими объёмами информации; методологии концептуального, логического и физического моделирования данных, адаптированные к распределённым системам (схемы «звезда» и «снежинка», денормализация под аналитические запросы, модель «Data Vault»); подходы к организации «Data Lake» и управлению метаданными, роль ETL/ELT-процессов в подготовке и оркестрации данных для аналитики и бизнес-отчётности, а также методы версионирования данных и построения витрин данных для обеспечения задач организационного управления и бизнес-процессов;

Уметь:

ПК-1.2/Ум1 Проектировать и создавать дизайн баз данных и хранилищ Big Data, разрабатывать логические и физические модели, оптимизированные для выполнения аналитических и поисковых запросов на сверхбольших наборах данных, проектировать слои «Data Lake» (сырой, очищенный, витринный), формировать ETL/ELT-пайплайны для трансформации и агрегации данных, а также управлять работами по модификации схем и реструктуризации хранилищ при эволюции бизнес-требований и росте объёмов данных в ИС организационного управления;

Владеть:

ПК-1.2/Нв1 Навыками проектирования хранилищ больших данных с применением CASE-средств и инструментов моделирования, а также навыками документирования структур данных, регламентов доступа и ETL-процессов при выполнении и управлении работами по сопровождению Big Data ИС.

ПК-1.3 Реализует процесс обеспечения и контроля качества в соответствии с регламентами организации в рамках выполнения работ и управления работами по созданию (модификации) и сопровождению ИС

Знать:

ПК-1.3/Зн1 Методологии обеспечения и контроля качества данных в системах больших данных, включая ключевые измерения качества, методы профилирования данных, выявления аномалий и проверки бизнес-правил в больших наборах; инструменты автоматизированного тестирования качества данных, практики CI/CD и DataOps для конвейеров обработки данных, подходы к мониторингу дрейфа данных и схем в ETL-процессах, а также регламенты организации в области управления качеством данных и обеспечения достоверности аналитической отчётности, формируемой ИС для задач организационного управления и бизнес-процессов;

Уметь:

ПК-1.3/Ум1 Применять методики обеспечения и контроля качества больших данных, выполнять профилирование и валидацию данных на всех этапах ETL/ELT-пайплайнов, выявлять и диагностировать дрейф данных и метаданных, оформлять отчёты о качестве данных и акты соответствия регламентам, а также управлять процессом устранения дефектов данных, обеспечивая надёжность и достоверность информационного фундамента ИС, автоматизирующих задачи менеджмента;

Владеть:

ПК-1.3/Нв1 Навыками реализации процессов обеспечения и контроля качества в экосистеме больших данных с использованием Python, а также навыками разработки регламентов, чек-листов и документации для аудита качества данных, ETL-процессов и аналитических выходов в рамках управления работами по созданию, модификации и индустриальному сопровождению Big Data ИС.

3. Место дисциплины в структуре ОП

Дисциплина (модуль) «Технологии больших данных» относится к формируемой участниками образовательных отношений части образовательной программы и изучается в семестре(ах): 5.

В процессе изучения дисциплины студент готовится к решению типов задач профессиональной деятельности, предусмотренных ФГОС ВО и образовательной программой.

Компетенция	Предшествующие дисциплины	Последующие дисциплины
ПК-1 - Способен выполнять работы и управлять работами по созданию (модификации) и сопровождению ИС, автоматизирующих задачи организационного управления и бизнес-процессы		

ПК-1.1 Разрабатывает архитектуру и прототип ИС в рамках выполнения работ и управления работами по созданию (модификации) и сопровождению ИС, автоматизирующих задачи организационного управления и бизнес-процессы	Архитектура интеллектуальных информационных систем, Веб-программирование, Интеллектуальные информационные системы в менеджменте, Информационно-коммуникационные технологии в профессиональной деятельности, Нейросетевые технологии в социальных медиа, Облачные ресурсы для разработки интеллектуальных сервисов, Облачные технологии и сервисы, Производственная практика: технологическая (проектно-технологическая) практика, Разработка распределенных приложений, Разработка цифровых сервисов на основе искусственного интеллекта, Системы машинного обучения в управлении, Учебная практика: ознакомительная практика, Цифровой дизайн, инфографика и визуализация данных в управлении	Автоматизированное тестирование программного обеспечения, Блокчейн и его приложения, Выполнение и защита выпускной квалификационной работы, Интеллектуальные информационные системы в менеджменте, Корпоративные интеллектуальные системы, Методы и средства проектирования цифровых сервисов в управлении, Проектирование и разработка экспертных интеллектуальных систем, Производственная практика: преддипломная практика, Производственная практика: технологическая (проектно-технологическая) практика, Разработка цифровых сервисов на основе искусственного интеллекта, Системы машинного обучения в управлении, Современные технологии разработки приложений, Технологии блокчейн, Управление качеством разработки приложений, Цифровой дизайн, инфографика и визуализация данных в управлении
ПК-1.2 Проектирует, создает дизайн и разрабатывает базы данных ИС в рамках выполнения работ и управления работами по созданию (модификации) и сопровождению ИС	Архитектура интеллектуальных информационных систем, Веб-программирование, Интеллектуальные информационные системы в менеджменте, Информационно-коммуникационные технологии в профессиональной деятельности, Нейросетевые технологии в социальных медиа, Облачные ресурсы для разработки интеллектуальных сервисов, Облачные технологии и сервисы, Производственная практика: технологическая (проектно-технологическая) практика, Разработка распределенных приложений, Разработка цифровых сервисов на основе искусственного интеллекта, Системы машинного обучения в управлении, Учебная практика: ознакомительная практика, Цифровой дизайн, инфографика и визуализация данных в управлении	Автоматизированное тестирование программного обеспечения, Блокчейн и его приложения, Выполнение и защита выпускной квалификационной работы, Интеллектуальные информационные системы в менеджменте, Корпоративные интеллектуальные системы, Проектирование и разработка экспертных интеллектуальных систем, Производственная практика: преддипломная практика, Производственная практика: технологическая (проектно-технологическая) практика, Разработка цифровых сервисов на основе искусственного интеллекта, Системы машинного обучения в управлении, Современные технологии разработки приложений, Технологии блокчейн, Управление качеством разработки приложений, Цифровой дизайн, инфографика и визуализация данных в управлении

ПК-1.3 Реализует процесс обеспечения и контроля качества в соответствии с регламентами организации в рамках выполнения работ и управления работами по созданию (модификации) и сопровождению ИС	Архитектура интеллектуальных информационных систем, Веб-программирование, Интеллектуальные информационные системы в менеджменте, Информационно-коммуникационные технологии в профессиональной деятельности, Нейросетевые технологии в социальных медиа, Облачные ресурсы для разработки интеллектуальных сервисов, Облачные технологии и сервисы, Производственная практика: технологическая (проектно-технологическая) практика, Разработка распределенных приложений, Разработка цифровых сервисов на основе искусственного интеллекта, Системы машинного обучения в управлении, Учебная практика: ознакомительная практика, Цифровой дизайн, инфографика и визуализация данных в управлении	Автоматизированное тестирование программного обеспечения, Блокчейн и его приложения, Выполнение и защита выпускной квалификационной работы, Интеллектуальные информационные системы в менеджменте, Корпоративные интеллектуальные системы, Методы и средства проектирования цифровых сервисов в управлении, Проектирование и разработка экспертных интеллектуальных систем, Производственная практика: преддипломная практика, Производственная практика: технологическая (проектно-технологическая) практика, Разработка цифровых сервисов на основе искусственного интеллекта, Системы машинного обучения в управлении, Современные технологии разработки приложений, Технологии блокчейн, Управление качеством разработки приложений, Цифровой дизайн, инфографика и визуализация данных в управлении
---	--	--

4. Объем дисциплины (модуля) и виды учебной работы

Период обучения	Общая трудоемкость (часы)	Общая трудоемкость (ЗЕТ)	Контактная работа (часы, всего)	Лабораторные занятия (часы)	Лекционные занятия (часы)	Групповая контактная работа (часы)	Индивидуальная контактная работа (часы)	Самостоятельная работа (часы)	Промежуточная аттестация
Пятый семестр	144	4	60	28	32	2	0,3	47,7	Экзамен
Всего	144	4	60	28	32	2	0,3	47,7	34

5. Содержание дисциплины (модуля)

5.1. Разделы, темы дисциплины и виды занятий (часы промежуточной аттестации не указываются)

Наименование раздела, темы	Всего	Лабораторные занятия	Лекционные занятия	Самостоятельная работа

Раздел 1. Технологии больших данных	107,7	28	32	47,7
Тема 1.1. Архитектура и жизненный цикл информационных систем больших данных в задачах организационного управления	9,7		4	5,7
Тема 1.2. Прототипирование информационных систем больших данных с использованием Python	14	4	4	6
Тема 1.3. Проектирование и дизайн хранилищ больших данных: Data Lake, Data Lakehouse и NoSQL-системы	14	4	4	6
Тема 1.4. Моделирование данных для распределённых систем: многомерные схемы, Data Vault и управление метаданными	14	4	4	6
Тема 1.5. ETL/ELT-процессы и оркестрация конвейеров обработки больших данных	14	4	4	6
Тема 1.6. Поточковая и пакетная обработка данных и формирование аналитических витрин	14	4	4	6
Тема 1.7. Обеспечение и контроль качества данных в Big Data: профилирование, мониторинг дрейфа и практики DataOps	14	4	4	6
Тема 1.8. Управление работами по созданию, модификации и сопровождению Big Data ИС: регламенты, CI/CD и командная разработка	14	4	4	6

5.2. Контрольные мероприятия по дисциплине

Вид контроля	Форма контроля/Оценочное средство
Текущий контроль	Тестирование
Промежуточная аттестация	Экзамен

№ п/п	Наименование раздела	Вид контроля/ используемые оценочные материалы	
		Текущий	Промежут. аттестация
1	Технологии больших данных	Тестирование	Экзамен

6. Оценочные материалы текущего контроля

1. Технологии больших данных Тестирование

№ п/п	Содержание вопроса		Компетенция
		Правильный ответ (ключ ответа)	
1	Выберите один вариант ответа Какая архитектурная концепция объединяет гибкость «Data Lake» и строгую схему хранилищ данных, поддерживая ACID-транзакции и версионирование? А) Классическое хранилище данных Б) Data Lakehouse В) Федеративная база данных Г) Data Mart		ПК-1
	Ответ:	Б	
2	Выберите один вариант ответа Какой инструмент является распределённой платформой для приёма и передачи потоковых данных в реальном времени, широко используемой в Big Data ИС? А) Apache Spark Б) Apache Hive В) Apache Kafka Г) Apache Hadoop		ПК-1
	Ответ:	В	
3	Выберите один вариант ответа Какая библиотека Python предназначена для автоматизированного тестирования и контроля качества данных в конвейерах обработки? А) pandas Б) Great Expectations В) Matplotlib Г) Seaborn		ПК-1
	Ответ:	Б	
4	Выберите один вариант ответа Какое измерение качества данных оценивает долю записей, не содержащих пропусков в критически важных атрибутах? А) Точность Б) Своевременность В) Полнота Г) Уникальность		ПК-1
	Ответ:	В	
5	Выберите один вариант ответа Какой язык запросов используется в Apache Hive для анализа данных, хранящихся в HDFS? А) SQL Б) HiveQL В) SPARQL Г) Cypher		ПК-1
	Ответ:	Б	
6	Установите соответствие Установите соответствие между слоем Data Lake и его назначением: Слой «Data Lake»: а) Сырой (Bronze) б) Очищенный (Silver) в) Витринный (Gold) г) Лендинговая зона (Landing) Назначение: 1) Агрегированные, очищенные данные для аналитических запросов 2) Данные в исходном виде, без изменений, для аудита и предобработки 3) Данные после базовой очистки и нормализации, готовые к дальнейшему анализу 4) Область временного хранения входящих данных до перемещения в сырой слой		ПК-1
	Ответ:	а–2, б–3, в–1, г–4	

7	Установите соответствие Установите соответствие между типом NoSQL-хранилища и примером его использования в Big Data ИС: Тип NoSQL: а) Колоночное б) Документное в) Графовое г) Ключ-значение Характерный пример использования: 1) Хранение временных рядов IoT-датчиков 2) Хранение JSON-логов активности пользователей 3) Анализ цепочек поставок и связей контрагентов 4) Кэширование результатов аналитических запросов Ответ: а–1, б–2, в–3, г–4	ПК-1
8	Установите соответствие Установите соответствие между инструментом и его ролью в DataOps/CI/CD-конвейере: Инструмент: а) Apache Airflow б) GitLab CI в) dbt г) Great Expectations Роль в конвейере обработки больших данных: 1) Оркестрация и планирование ETL-джобов 2) Автоматизация сборки, тестирования и развертывания кода 3) Трансформация данных и проверка качества в хранилище 4) Профилирование и валидация данных на всех этапах Ответ: а–1, б–2, в–3, г–4	ПК-1
9	Установите правильную последовательность Установите правильную последовательность этапов классического ETL-процесса с контролем качества: А) Трансформация (очистка, агрегация, инжинерия признаков) Б) Загрузка в целевое хранилище В) Извлечение данных из источников Г) Валидация качества и полноты загруженных данных Ответ: В → А → Б → Г	ПК-1
10	Установите правильную последовательность Расположите этапы создания Big Data ИС в логическом порядке: А) Прототипирование системы на тестовом наборе данных Б) Сбор и анализ бизнес-требований к аналитике и отчётности В) Проектирование архитектуры Data Lake/Lakehouse Г) Внедрение, мониторинг и сопровождение системы Ответ: Б → В → А → Г	ПК-1
11	Дайте правильный ответ Назовите технологию, которая добавляет ACID-транзакции и версионирование данных поверх Data Lake на основе Apache Spark, позволяя реализовать модель Data Lakehouse. Ответ: Delta Lake	ПК-1
12	Дайте правильный ответ Как называется явление постепенного изменения статистических характеристик входных данных, способное привести к деградации качества аналитики и моделей в Big Data ИС? Ответ: Дрейф данных	ПК-1
13	Дайте правильный ответ Укажите название модели архитектуры хранилища, которая разделяет данные на концентраторы, связи и сателлиты для обеспечения гибкости и аудируемости в распределённых системах Ответ: Data Vault	ПК-1
14	Дайте правильный ответ Задача с числовым ответом. Система потоковой обработки принимает 5000 сообщений в секунду. Средний размер одного сообщения - 2 КБ. Каков объём данных в мегабайтах, накопленный за 10 минут? Ответ округлите до целого числа. Ответ: 5859 КБ	ПК-1
15	Дайте правильный ответ Какой практикой называется подход, объединяющий Agile, DevOps и инженерию данных для ускорения и повышения качества разработки конвейеров больших данных? Ответ: DataOps	ПК-1

7. Оценочные материалы промежуточной аттестации

Экзамен пятый семестр

№ п/п	Содержание вопроса		Компетенция
		Правильный ответ (ключ ответа)	
1	Дайте развернутый ответ Охарактеризуйте современные архитектурные концепции информационных систем больших данных - Data Lake и Data Lakehouse. Опишите их компоненты, достоинства, ограничения и жизненный цикл создания подобных систем в контексте автоматизации организационного управления.		ПК-1
	Ответ:	Data Lake - это архитектурный подход, предполагающий сохранение всех данных организации в сыром виде в едином масштабируемом хранилище (чаще всего на базе HDFS или объектных хранилищ S3-совместимых систем). Ключевой принцип - «схема при чтении» (schema-on-read): структура и смысл данных определяются в момент их использования, а не при загрузке. Это обеспечивает гибкость, позволяет хранить структурированные, полуструктурированные и неструктурированные данные (логи, изображения, JSON) и снижает затраты на предварительную подготовку. Компоненты Data Lake включают зону посадки (landing), сырой слой (bronze), очищенный (silver) и витринный (gold). Однако классический Data Lake страдает от проблем с качеством данных, отсутствия транзакционной поддержки и сложности управления версиями, что может превратить его в «болото данных» (data swamp). Data Lakehouse - эволюция концепции, объединяющая гибкость Data Lake с надёжностью хранилищ данных. Ключевые характеристики: ACID-транзакции, поддержка SQL, версионирование, управление схемой и индексация - и всё это поверх дешёвого распределённого хранения. Реализации (Delta Lake, Apache Iceberg, Apache Hudi) добавляют слой метаданных, который отслеживает версии файлов, позволяет выполнять конкурентные запросы и обеспечивает временные запросы (time travel). Lakehouse подходит для задач организационного управления, где требуется как аналитика больших объёмов, так и надёжность отчётности. Жизненный цикл создания Big Data ИС начинается с анализа бизнес-требований к аналитике и отчётности. Далее проектируется архитектура с выбором слоёв хранения, инструментов интеграции (Kafka для потоков, Spark для обработки) и BI-инструментов. Создаётся прототип на тестовых данных, оценивается масштабируемость. После внедрения система сопровождается: мониторятся потоки данных, контролируется качество, выполняется реструктуризация при росте объёмов. Управление работами по модификации опирается на регламенты версионирования и командные практики CI/CD.	
2	Дайте развернутый ответ Опишите принципы и инструменты быстрого прототипирования систем больших данных с использованием Python, PySpark и облачных сервисов. Какие задачи позволяет решить прототип и как он встраивается в процесс создания Big Data ИС?		ПК-1
	Ответ:	Быстрое прототипирование в области больших данных преследует цель в сжатые сроки проверить гипотезы о ценности данных, работоспособности архитектурных решений и ожидаемой производительности, не разворачивая полноценную промышленную среду. Основным инструментом выступает Python, дополненный библиотеками PySpark, pandas и средствами визуализации. PySpark предоставляет унифицированный API для пакетной и потоковой обработки, позволяя тот же код, что будет работать на кластере, отлаживать локально на небольших выборках. Облачные сервисы (AWS EMR, Google Dataproc, Azure Synapse) дают возможность за минуты развернуть управляемые Spark-кластеры, подключить к объектным хранилищам и начать обработку данных без капитальных затрат на инфраструктуру. Процесс прототипирования включает: сбор небольшого репрезентативного набора данных из источников; разведочный анализ (EDA) в Jupyter Notebook с визуализацией; написание пайплайна трансформации на PySpark, который читает сырые данные, очищает, агрегирует и формирует витрину; проверку корректности результатов и оценку времени выполнения. Код прототипа часто оформляется как набор ноутбуков или скриптов, которые демонстрируются бизнес-заказчикам. Прототип решает задачи подтверждения возможности интеграции источников, проверки гипотез о необходимых трансформациях, а также позволяет уточнить нефункциональные требования - объёмы хранения, пропускную способность. После успешной демонстрации архитектура дорабатывается, код рефакторится в модули и пайплайны, настраивается оркестрация (Airflow). Умение быстро создавать и демонстрировать работающий прототип - ключевая компетенция при управлении созданием и модификацией ИС, так как снижает риски недопонимания требований и дорогостоящих ошибок на поздних стадиях.	
3	Дайте развернутый ответ Проведите сравнительный анализ типов NoSQL-хранилищ (колоночные, документные, графовые, ключ-значение) применительно к задачам больших данных в организационном управлении. Поясните критерии выбора хранилища при проектировании Big Data ИС.		ПК-1

	Ответ: Проектирование хранилищ для Big Data требует учёта модели данных, характера запросов и требований к масштабируемости. Реляционные аналитические СУБД (Greenplum, ClickHouse) эффективны для структурированных данных с заранее известной схемой и сложных SQL-запросов, но их масштабирование часто сложнее и дороже, чем у NoSQL. Колоночные хранилища (Apache Cassandra, HBase, ClickHouse как колоночная СУБД) оптимизированы для операций чтения и записи больших объёмов с преобладанием вставок. Cassandra подходит для сбора временных рядов и логов благодаря линейной масштабируемости и отсутствию единой точки отказа. Документные хранилища (MongoDB) хранят данные в формате JSON, что удобно для слабоструктурированной информации: профили пользователей, параметры конфигураций, результаты сбора веб-аналитики. Гибкая схема позволяет быстро адаптироваться к изменениям источников данных, что востребовано при модификации ИС. Графовые базы (Neo4j) незаменимы, когда ценность представляют связи: организационные иерархии, цепочки поставок, графы взаимодействия контрагентов. Они позволяют выполнять навигационные запросы (например, «найти кратчайший путь влияния») гораздо эффективнее, чем многократные JOIN в SQL. Хранилища типа «ключ-значение» (Redis, DynamoDB) обеспечивают минимальную задержку и часто применяются для кэширования, хранения сессий, справочников, а также как быстрый слой для результатов аналитики. При проектировании Big Data ИС выбор осуществляется на основе требований к консистентности (CAP-теорема), паттернам доступа (чтение/запись), требуемой гибкости схемы и сложности запросов. Современная архитектура часто использует полиглотную модель: например, сырые данные хранятся в Data Lake на S3, очищенные и агрегированные загружаются в колоночную СУБД для отчётности, а поисковые сценарии обслуживаются графовой и документной базами. Умение обоснованно выбирать и проектировать такие комбинации является частью компетенции по созданию и модификации ИС.	
4	Дайте развернутый ответ Раскройте методологии моделирования данных, адаптированные к распределённым системам больших данных: многомерные схемы, Data Vault, подходы к денормализации. Опишите их роль в построении хранилищ и управлении метаданными. Ответ: Моделирование данных для Big Data отличается от традиционного OLTP-проектирования необходимостью оптимизации под аналитические запросы на огромных объёмах. Многомерные схемы «звезда» и «снежинка» остаются основой для построения витрин данных. В схеме «звезда» центральная таблица фактов (транзакции, события) окружена денормализованными таблицами измерений (время, продукт, клиент). Это минимизирует количество JOIN-ов, что критично для быстрого действия распределённых движков типа Hive и Spark SQL. Схема «снежинка» нормализует измерения, экономя место, но увеличивая сложность запросов. Для Data Lakehouse эти схемы реализуются прямо на Parquet-файлах, а оптимизацию обеспечивают Z-ordering и секционирование. Модель Data Vault специально создана для масштабируемых хранилищ с высокой изменчивостью источников. Она разделяет данные на три типа сущностей: концентраторы (хабы) - хранят бизнес-ключи (идентификатор клиента); связи (линки) - описывают отношения между ключами; сателлиты - хранят атрибуты и историю их изменений с временными метками. Такой подход обеспечивает гибкость: добавление нового источника данных не ломает существующую схему, а требует лишь новых сателлитов. Data Vault хорошо сочетается с концепцией Data Lake, поддерживая аудит и полную восстанавливаемость. Денормализация - осознанное нарушение нормальных форм ради производительности. В Big Data это стандартная практика: предварительно вычисленные агрегаты, вложенные JSON-поля, дублирование атрибутов для исключения JOIN-ов. Управление метаданными (Apache Atlas, AWS Glue Data Catalog) связывает модели с реальными файлами и таблицами. Каталоги хранят схемы, статистику, lineage (происхождение данных), что критично для поиска данных в Data Lake и контроля качества. Навык владения этими методологиями позволяет проектировать хранилища, способные эффективно эволюционировать вместе с бизнес-требованиями, обеспечивая надёжную основу для аналитической отчётности ИС организационного управления.	ПК-1
5	Дайте развернутый ответ Опишите устройство ETL/ELT-процессов в экосистеме больших данных: основные фазы, инструменты оркестрации и методы обеспечения качества на каждом этапе. Приведите типовую архитектуру конвейера обработки данных.	ПК-1

	Ответ: ETL (Extract, Transform, Load) - последовательность извлечения данных из источников, их трансформации и загрузки в целевое хранилище. В контексте больших данных акцент смещается на ELT: данные сначала загружаются в Data Lake в сыром виде, а трансформация выполняется позже с использованием мощностей распределённых движков (Spark, Hive). Это даёт гибкость, так как исходные данные сохраняются и могут быть переобработаны при изменении бизнес-логики. Фазы процесса: 1. Извлечение - сбор данных из реляционных БД (через Sqoop, JDBC), потоковых источников (Kafka), API, файлов. Для полноты и своевременности применяются инкрементальные стратегии. 2. Первичная загрузка в сырой слой Data Lake (бронза). 3. Трансформация: очистка (удаление дубликатов, обработка NULL), нормализация, инженерия признаков, агрегация. Выполняется с помощью PySpark/Spark SQL, результаты сохраняются в очищенном (серебряном) слое в форматах Parquet/ORC. 4. Формирование витрин (золотой слой) - денормализованные таблицы, оптимизированные под BI-запросы. 5. Оркестрация всего процесса выполняется планировщиками (Apache Airflow, Dagster, Prefect), которые запускают джобы по расписанию или событиям, обрабатывают сбои и обеспечивают отслеживание lineage. Обеспечение качества встраивается в конвейер: после извлечения проверяется полнота (число записей), на этапе трансформации - корректность агрегатов, на выходе витрин - соответствие бизнес-правилам (например, сумма балансов равна итогу). Инструменты вроде Great Expectations позволяют декларативно описать ожидания к данным и автоматически формировать отчёт о качестве. При обнаружении аномалий конвейер может остановиться или отправить алерт. Типовая архитектура конвейера: источники → Kafka/Airflow-сенсоры → Spark Structured Streaming/пакетный Spark → Delta Lake (хранение) → dbt для трансформации и тестов → BI-инструмент. Такой подход обеспечивает надёжный и контролируемый поток данных для принятия управленческих решений.	
6	Дайте развернутый ответ Раскройте технологии потоковой и пакетной обработки данных в Big Data ИС. Опишите роль Apache Kafka и Apache Spark Streaming в построении систем реального времени для мониторинга бизнес-процессов. Ответ: Пакетная обработка (batch) оперирует накопленными за период данными и характерна для регулярной аналитики: суточные отчёты, пересчёт моделей. Потоковая обработка (stream) работает с непрерывным потоком событий, обеспечивая задержки от миллисекунд до секунд, что необходимо для мониторинга, алертинга и реактивных действий. В архитектуре больших данных эти подходы часто комбинируются (лямбда- и каппа-архитектуры). Apache Kafka - де-факто стандарт для приёма и буферизации потоков. Это распределённый журнал сообщений (message broker), публикующий данные по топикам и обеспечивающий долговременное хранение. В ИС организационного управления Kafka собирает события из различных бизнес-приложений: факты продаж, изменения статусов заказов, клики на портале, показания датчиков. Высокая пропускная способность и отказоустойчивость гарантируют, что ни одно событие не будет потеряно. Для обработки этих потоков используется Spark Streaming (модуль Apache Spark) или более современный Spark Structured Streaming, который позволяет писать потоковые запросы так же, как пакетные. Он поддерживает оконные агрегации (подсчёт событий за последние 5 минут), соединение потока со статическими справочниками, запись результатов в файлы, Kafka или базы данных. Например, система мониторинга бизнес-процессов в реальном времени читает из Kafka поток транзакций, агрегирует их по регионам и сравнивает с плановыми показателями, отправляя оповещение при отклонениях. При создании прототипа такой системы разработчик на Python с PySpark может в Jupyter-ноутбуке смоделировать поток с чтением из Kafka, написать агрегации и вывести дашборд в Streamlit, демонстрируя работу в реальном времени. Внедрение потоковой обработки существенно повышает ценность ИС для менеджмента, обеспечивая переход от реактивной отчётности к проактивному управлению.	ПК-1
7	Дайте развернутый ответ Систематизируйте ключевые измерения качества данных в системах больших данных и опишите методологию профилирования и автоматизированной валидации с использованием Great Expectations. Как это вписывается в регламенты контроля качества организации?	ПК-1

	Ответ: Качество данных - фундамент доверия к аналитике. Основные измерения: полнота (completeness) - отсутствие пропусков в критических атрибутах; точность (accuracy) - соответствие реальному миру; согласованность (consistency) - непротиворечивость между связанными наборами; своевременность (timeliness) - доступность данных в требуемый момент; уникальность (uniqueness) - отсутствие дубликатов; валидность (validity) - соответствие форматам и диапазонам. В Big Data проблемы качества усугубляются объёмом и разнообразием источников, поэтому проверки должны быть автоматизированы. Профилирование - начальный этап: сбор статистик (количество строк, число уникальных значений, min/max, распределения) и выявление аномалий. Инструменты: Apache Griffin, Deequ (Spark-библиотека) и Great Expectations (GE). GE особенно популярен благодаря декларативному подходу: ожидания (expectations) описываются в виде «ожидаю, что столбец 'transaction_date' не содержит null», «ожидаю, что значения 'amount' больше 0». Эти ожидания группируются в наборы и запускаются как автоматические тесты в пайплайне. Результатом является «документ качества данных» - JSON-отчёт с числом успешных и проваленных проверок. В контексте регламентов организации GE интегрируется в CI/CD-конвейер: при каждом обновлении данных джоба Airflow запускает Spark-шаг трансформации, а затем проверочный шаг на базе GE. Если процент неудачных ожиданий превышает порог, конвейер останавливается, отправляется уведомление владельцу данных. Такой подход документирует контроль качества, создавая аудиторский след, и позволяет реализовать процесс управления качеством, соответствующий корпоративным стандартам. Умение настраивать и интерпретировать такие проверки является частью компетенций по сопровождению Big Data ИС, гарантируя достоверность отчётности для организационного управления.	
8	Дайте развернутый ответ Что такое практика DataOps и как она реализуется с помощью CI/CD для конвейеров обработки больших данных? Опишите роль оркестраторов, dbt и систем контроля версий в обеспечении качества и скорости поставки аналитики. Ответ: DataOps - это методология, адаптирующая принципы DevOps и Agile к инженерии данных. Её цель - сократить время от гипотезы до развёртывания аналитического продукта, повысить качество данных и воспроизводимость результатов. Ключевые принципы: автоматизация пайплайнов, непрерывная интеграция и доставка (CI/CD), мониторинг, командная работа с кодом и данными. Реализация CI/CD для больших данных отличается от классической разработки ПО тем, что помимо кода необходимо версионировать схемы данных, зависимости и сами конвейеры. Конвейер CI/CD может выглядеть так: код трансформаций (PySpark/SQL) и определения качества (Expectations) хранятся в Git. При каждом коммите GitLab CI/GitHub Actions запускает легковесное окружение (Docker-контейнер с Spark), выполняет dbt-компиляцию моделей, прогоняет тесты на синтетических данных и проверяет метаданные. Затем изменения проходят код-ревью. После слияния в основную ветку Airflow автоматически забирает новый DAG и разворачивает обновлённый пайплайн на стейджинге, где выполняются более тяжёлые проверки на реальных выборках. Финальный промоушен в продакшн контролируется вручную или по метрикам качества. Особую роль играют dbt (data build tool) и оркестраторы. dbt позволяет описывать трансформации как SQL-селекты с поддержкой шаблонов (Jinja), зависимостями между моделями и встроенным тестированием (уникальность, ссылочная целостность). Airflow/Dagster организует выполнение задач в нужном порядке и отслеживает сбои. Такой стек обеспечивает модульность, повторяемость и скорость разработки. DataOps требует также культуры документирования и мониторинга. Владение этими практиками позволяет управлять созданием и модификацией ИС, быстро реагировать на изменения бизнес-требований и поддерживать высокое доверие к данным, что является критичным для принятия управленческих решений.	ПК-1
9	Дайте развернутый ответ Опишите процесс управления работами по созданию, модификации и сопровождению Big Data ИС. Какие регламенты и командные роли необходимы для успешной эксплуатации системы, автоматизирующей задачи организационного управления?	ПК-1

	Ответ: Управление работами в проектах больших данных требует координации разнопрофильных специалистов (data-инженеров, аналитиков, администраторов, менеджеров) и следования установленным регламентам. Жизненный цикл начинается с инициации, где формулируются бизнес-цели и критерии успеха. Далее проектируется архитектура, создаётся прототип, выполняется промышленная разработка, внедрение и эксплуатация. Регламенты организации должны охватывать: • Управление версиями - не только кода (Git), но и данных (Delta Lake, DVC) и схем; все изменения проходят через процесс код-ревью. • Документирование - архитектурные решения (ADR), модели данных, lineage, инструкции по эксплуатации. • Управление инцидентами - порядок реагирования на сбои данных, падения пайплайнов, деградацию производительности. • Контроль качества - обязательные этапы тестирования и валидации перед выкаткой в продуктив. • Безопасность и доступ - регламенты разграничения прав к данным, аудит действий. Командная работа строится по гибким методологиям (Scrum/Kanban). Роль Product Owner транслирует бизнес-требования. Data-инженеры разрабатывают и поддерживают ETL/ELT. Data Steward отвечает за качество и семантику данных. DevOps/MLOps-инженер обеспечивает инфраструктуру, CI/CD и мониторинг. Менеджер проекта координирует сроки и коммуникации. Сопровождение ИС включает регулярный мониторинг объёмов и качества данных, своевременное обновление моделей данных при изменении источников, оптимизацию запросов и плановое наращивание мощностей. Процесс модификации запускается при появлении новых бизнес-потребностей или слиянии данных из новых систем и проходит тот же цикл: анализ, дизайн, тестирование, развёртывание. Управляющий работами должен обеспечить прозрачность состояния системы, соблюдение SLA по доступности и достоверности отчётности, что в конечном счёте определяет ценность Big Data ИС для организационного управления.	
10	Дайте развернутый ответ Раскройте методы версионирования данных и управления метаданными в Data Lake на примере Delta Lake и Apache Atlas. Почему эти компоненты критичны для сопровождения Big Data ИС? Ответ: Версионирование данных и управление метаданными решают ключевые проблемы Data Lake: «болото данных», невозможность воспроизводимости аналитики и сложность поиска нужных наборов. Delta Lake - это слой с открытым исходным кодом, надстроенный над Data Lake (обычно на базе Parquet-файлов), который добавляет ACID-транзакции, управление версиями и индексацию. Он хранит журнал транзакций (_delta_log), в который записываются все операции вставки, обновления и удаления. Это позволяет: читать согласованные снимки данных на любой момент (time travel), откатывать изменения, предотвращать конфликты при одновременной записи. Для сопровождения это значит, что любое изменение витрин отслеживаемо, можно восстановить состояние до ошибочного обновления, а аналитики всегда видят непротиворечивые данные. Apache Atlas - система управления метаданными и governance. Она автоматически собирает технические метаданные (схемы таблиц Hive, местоположение файлов, тэги) и строит граф lineage, показывающий, как данные движутся от источников к отчётам. Бизнес-пользователи могут добавлять глоссарий бизнес-терминов (например, «выручка» связывается с конкретной колонкой в нескольких БД). Atlas интегрируется со стеком больших данных (Spark, Hive, Kafka) через хуки, а также предоставляет REST API и UI для поиска. Эти компоненты критичны для сопровождения, так как обеспечивают прозрачность и доверие. При обнаружении ошибки в отчёте lineage позволяет быстро найти её источник. Аудит, требуемый регламентами, легко выполняется по журналу версий Delta и метаданным Atlas. Версионирование данных также открывает возможности для автоматизации: при появлении новой порции данных CI/CD-пайплайн может прогонять тесты на качество, а в случае проблем откатывать витрину к предыдущей стабильной версии, минимизируя влияние на бизнес-отчётность. Владение этими инструментами - неотъемлемая часть компетенции по управлению работами по созданию и сопровождению Big Data ИС.	ПК-1

7.1. Уровни овладения

Компетенция: ПК-1 Способен выполнять работы и управлять работами по созданию (модификации) и сопровождению ИС, автоматизирующих задачи организационного управления и бизнес-процессы.

Индикатор достижения компетенции: ПК-1.1 Разрабатывает архитектуру и прототип ИС в рамках выполнения работ и управления работами по созданию (модификации) и сопровождению ИС, автоматизирующих задачи организационного управления и бизнес-процессы.

Уровень	Характеристика	Оценка в баллах
---------	----------------	-----------------

Повышенный	Достигнуто полное овладение знаниями, умениями и навыками. Студент свободно владеет терминологией, умеет применять теоретические знания в различных ситуациях для решения поставленных задач.	81-100
Базовый	Достигнуто достаточное овладение знаниями, умениями и навыками. Студент уверенно владеет терминологией, умеет применять теоретические знания в различных ситуациях для решения поставленных задач.	61-80
Пороговый	Достигнуто овладение минимально необходимыми знаниями, умениями и навыками. Студент владеет основной терминологией, умеет применять теоретические знания для решения поставленных задач в стандартных ситуациях.	41-60
Ниже порогового	Компетенция не освоена	0-40

Индикатор достижения компетенции: ПК-1.2 Проектирует, создает дизайн и разрабатывает базы данных ИС в рамках выполнения работ и управления работами по созданию (модификации) и сопровождению ИС.

Уровень	Характеристика	Оценка в баллах
Повышенный	Достигнуто полное овладение знаниями, умениями и навыками. Студент свободно владеет терминологией, умеет применять теоретические знания в различных ситуациях для решения поставленных задач.	81-100
Базовый	Достигнуто достаточное овладение знаниями, умениями и навыками. Студент уверенно владеет терминологией, умеет применять теоретические знания в различных ситуациях для решения поставленных задач.	61-80
Пороговый	Достигнуто овладение минимально необходимыми знаниями, умениями и навыками. Студент владеет основной терминологией, умеет применять теоретические знания для решения поставленных задач в стандартных ситуациях.	41-60
Ниже порогового	Компетенция не освоена	0-40

Индикатор достижения компетенции: ПК-1.3 Реализует процесс обеспечения и контроля качества в соответствии с регламентами организации в рамках выполнения работ и управления работами по созданию (модификации) и сопровождению ИС.

Уровень	Характеристика	Оценка в баллах
Повышенный	Достигнуто полное овладение знаниями, умениями и навыками. Студент свободно владеет терминологией, умеет применять теоретические знания в различных ситуациях для решения поставленных задач.	81-100
Базовый	Достигнуто достаточное овладение знаниями, умениями и навыками. Студент уверенно владеет терминологией, умеет применять теоретические знания в различных ситуациях для решения поставленных задач.	61-80

Пороговый	Достигнуто овладение минимально необходимыми знаниями, умениями и навыками. Студент владеет основной терминологией, умеет применять теоретические знания для решения поставленных задач в стандартных ситуациях.	41-60
Ниже порогового	Компетенция не освоена	0-40

8. Материально-техническое и учебно-методическое обеспечение дисциплины

8.1. Перечень основной и дополнительной учебной литературы

Основная литература

1. Воронов, М. В. Системы искусственного интеллекта: учебник и практикум для вузов / М. В. Воронов, В. И. Пименов, И. А. Небаев. - 2-е изд. - Москва: Юрайт, 2026. - 268 с - 978-5-534-17032-0. - Текст: электронный // ИКО Юрайт: [сайт]. - URL: <https://urait.ru/bcode/588642> (дата обращения: 21.05.2026). - Режим доступа: по подписке
2. Новиков, Ф. А. Символический искусственный интеллект: математические основы представления знаний: учебник для вузов / Ф. А. Новиков. - Москва: Юрайт, 2026. - 278 с - 978-5-534-00734-3. - Текст: электронный // ИКО Юрайт: [сайт]. - URL: <https://urait.ru/bcode/584114> (дата обращения: 21.05.2026). - Режим доступа: по подписке

Дополнительная литература

1. Трофимов, В. В. Интеллектуальные системы поддержки принятия решений. Цифровая трансформация, искусственный интеллект: учебник для вузов / В. В. Трофимов, Е. В. Трофимова. - Москва: Юрайт, 2026. - 199 с - 978-5-534-21777-3. - Текст: электронный // ИКО Юрайт: [сайт]. - URL: <https://urait.ru/bcode/590642> (дата обращения: 21.05.2026). - Режим доступа: по подписке

8.2. Профессиональные базы данных и ресурсы «Интернет», к которым обеспечивается доступ обучающихся

Профессиональные базы данных

Не используются.

Ресурсы «Интернет»

1. <https://cyberleninka.ru/> - Научная электронная библиотека «КиберЛенинка»
2. <https://stackoverflow.com> - Stack Overflow

8.3. Программное обеспечение и информационно-справочные системы, используемые при осуществлении образовательного процесса по дисциплине

Перечень программного обеспечения

(обновление производится по мере появления новых версий программы)

1. Microsoft Word;
2. Microsoft Excel;
3. Mozilla Firefox;
4. Python 3 (GPL-compatible);

Перечень информационно-справочных систем

(обновление выполняется еженедельно)

Не используются.

8.4. Специальные помещения, лаборатории и лабораторное оборудование

Учебные аудитории для проведения занятий лекционного типа	Комплекты ученической мебели Мультимедийный проектор Доска Экран
Учебные аудитории для проведения практических занятий (занятий семинарского типа)	Комплекты ученической мебели Мультимедийный проектор Доска Экран Компьютеры с выходом в сеть «Интернет» и ЭИОС СПб
Учебные аудитории для групповых и индивидуальных консультаций	Комплекты ученической мебели Мультимедийный проектор Доска Экран Компьютеры с выходом в сеть «Интернет» и ЭИОС СПб
Учебные аудитории для текущего контроля и промежуточной аттестации	Комплекты ученической мебели Мультимедийный проектор Доска Экран Компьютеры с выходом в сеть «Интернет» и ЭИОС СПб
Помещения для самостоятельной работы	Комплекты ученической мебели Мультимедийный проектор Доска Экран Компьютеры с выходом в сеть «Интернет» и ЭИОС СПб
Помещения для хранения и профилактического обслуживания оборудования	Комплекты специализированной мебели для хранения