МИНОБРНАУКИ РОССИИ

ВЛАДИВОСТОКСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

НАУЧНО-ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР "ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ"

Рабочая программа дисциплины (модуля)

ОСНОВЫ DATA ENGINEERING

Направление и направленность (профиль)
09.04.03 Прикладная информатика. Искусственный интеллект и машинное обучение в управлении и принятии решений

Год набора на ОПОП
2024

Форма обучения
очная

Владивосток 2025

Рабочая программа дисциплины (модуля) «Основы Data Engineering» составлена в соответствии с требованиями ФГОС ВО по направлению подготовки 09.04.03 Прикладная информатика (утв. приказом Минобрнауки России от 19.09.2017г. №916) и Порядком организации и осуществления образовательной деятельности по образовательным программам высшего образования – программам бакалавриата, программам специалитета, программам магистратуры (утв. приказом Минобрнауки России от 06.04.2021 г. N245).

Составитель(и):

Шахгельдян К.И., доктор технических наук, профессор, Научно-образовательный центр "Искусственный интеллект", carina.shahgeldyan@vvsu.ru

Утверждена на заседании научно-образовательный центр "искусственный интеллект" от 05.06.2025 , протокол № 6

СОГЛАСОВАНО:

Заведующий кафедрой (разработчика)

Кригер А.Б.

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат	1582918206
Номер транзакции	0000000000DC4DD5
Владелец	Кригер А.Б.

1 Цель, планируемые результаты обучения по дисциплине (модулю)

Целью дисциплины является знакомство с технологиями больших данных и получение начальных навыков развертывания инфраструктуры Big data.

Задачи освоения дисциплины:

1. Познакомиться с основными открытыми технологиями больших данных
2. Получить начальные навыки развертывания и сопровождения инфраструктуры Big Data

Планируемыми результатами обучения по дисциплине (модулю), являются знания, умения, навыки. Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы, представлен в таблице 1.

Таблица 1 – Компетенции, формируемые в результате изучения дисциплины (модуля)

Название ОПОП ВО, сокращенное	Код и формулировка компетенции	Код и формулировка индикатора достижения компетенции	Результаты обучения по дисциплине	
			Код результата	Формулировка результата
09.04.03 «Прикладная информатика» (М-ПИ)				

В процессе освоения дисциплины решаются задачи воспитания гармонично развитой, патриотичной и социально ответственной личности на основе традиционных российских духовно-нравственных и культурно-исторических ценностей, представленные в таблице 1.2.

Таблица 1.2 – Целевые ориентиры воспитания

Воспитательные задачи	**Формирование ценностей**	**Целевые ориентиры**
Формирование научного мировоззрения и культуры мышления		
Формирование осознания ценности научного мировоззрения и критического мышления	Гуманизм	Системное мышление
Формирование коммуникативных навыков и культуры общения		
Формирование навыков публичного выступления и презентации своих идей	Взаимопомощь и взаимоуважение	Умение работать в команде и взаимопомощь

2 Место дисциплины (модуля) в структуре ОПОП

Дисциплина «Технологии Big Data» относится обязательной части Блока 1 Дисциплины (модули).

3. Объем дисциплины (модуля)

Объем дисциплины (модуля) в зачетных единицах с указанием количества академических часов, выделенных на контактную работу с обучающимися (по видам учебных занятий) и на самостоятельную работу, приведен в таблице 2.

Таблица 2 – Общая трудоемкость дисциплины

Название ОПОП ВО	Форма обучения	Часть УП	Семестр (ОФО) или курс (ЗФО, ОЗФО)	Трудоемкость	Объем контактной работы (час)						СРС	Форма аттестации
				(З.Е.)	Всего	Аудиторная			Внеаудиторная			
						лек.	прак.	лаб.	ПА	КСР		
09.04.03 Прикладная информатика	ОФО	М01.Б	2	4	37	0	36	0	1	0	107	ДЗ

4 Структура и содержание дисциплины (модуля)

4.1 Структура дисциплины (модуля) для ОФО

Тематический план, отражающий содержание дисциплины (перечень разделов и тем), структурированное по видам учебных занятий с указанием их объемов в соответствии с учебным планом, приведен в таблице 3.1

Таблица 3.1 – Разделы дисциплины (модуля), виды учебной деятельности и формы текущего контроля для ОФО

№	Название темы	Код результата обучения	Кол-во часов, отведенное на				Форма текущего контроля
			Лек	Практ	Лаб	СРС	
1	Генерация данных	РД1	0	6	0	11	Собеседование
2	Инфраструктура больших данных	РД2, РД4, РД5	0	12	0	30	Собеседование
3	MapReduce	РД3, РД4	0	6	0	30	Экзамен/собеседование
4	Технологии Big Data	РД5	0	12	0	36	Экзамен/собеседование
Итого по таблице			**0**	**36**	**0**	**107**	

4.2 Содержание разделов и тем дисциплины (модуля) для ОФО

Тема 1 Генерация данных.

Содержание темы: Источники больших данных: поисковые машины, социальные сети, банковские транзакции, телеком, биоинформатика, Интернет вещей. Характеристики больших данных. Применение больших данных в жизни человека.

Формы и методы проведения занятий по теме, применяемые образовательные технологии: Практическое занятие.

Виды самостоятельной подготовки студентов по теме: .

Тема 2 Инфраструктура больших данных.

Содержание темы: Распределенные архитектуры больших данных. Параллельные и распределенные вычисления. Grid. Распределенные файловые системы. Файловая система Google. Основные компоненты распределенной файловой системы. Восстановление данных в распределенной файловой системе.

Формы и методы проведения занятий по теме, применяемые образовательные технологии: Практическое занятие.

Виды самостоятельной подготовки студентов по теме: .

Тема 3 MapReduce.

Содержание темы: Модель распределенных вычислений в вычислительных кластерах.

Формы и методы проведения занятий по теме, применяемые образовательные технологии: Практическое занятие.

Виды самостоятельной подготовки студентов по теме: .

Тема 4 Технологии Big Data.

Содержание темы: Технологии распределенных вычислений и хранения данных: Hadoop – технология распределенных вычисления на основе модели Map Reduce. HDFS – технология распределенной файловой системы Hadoop. Решение на базе Hadoop: Cloudera. Технология параллельной обработки данных Apache nifi. Диспетчеризация сообщений Apache Kafka. Системы хранения AsterixDB, HP Vertica, Impala, Neo4j, Redis, SparkSQL.

Формы и методы проведения занятий по теме, применяемые образовательные технологии: Практическое занятие.

Виды самостоятельной подготовки студентов по теме: .

5 Методические указания для обучающихся по изучению и реализации дисциплины (модуля)

5.1 Методические рекомендации обучающимся по изучению дисциплины и по обеспечению самостоятельной работы

Методические рекомендации по организации самостоятельной работы

В ходе изучения дисциплины студенты должны посещать аудиторные занятия (практические занятия, консультации). Особое место в овладении частью тем данной дисциплины отводится самостоятельной работе, при этом во время аудиторных занятий могут быть рассмотрены и проработаны наиболее важные и трудные вопросы по той или иной теме дисциплины, а применение уже освоенные навыков в смежных технологиях вынесены на самостоятельное обучение.

В соответствии с учебным планом направления подготовки процесс изучения дисциплины предусматривает проведение практических занятий, консультаций, а также самостоятельную работу студентов.

Ниже перечислены предназначенные для самостоятельного изучения студентами те вопросы, которые во время проведения аудиторных занятий изучаются недостаточно или изучение которых носит обзорный характер.

Перечень и тематика самостоятельных работ студентов по дисциплине

1. Использование базы MongoDB
2. Инструменты интеграции Splunk и Datameer

5.2 Особенности организации обучения для лиц с ограниченными возможностями здоровья и инвалидов

При необходимости обучающимся из числа лиц с ограниченными возможностями здоровья и инвалидов (по заявлению обучающегося) предоставляется учебная информация в доступных формах с учетом их индивидуальных психофизических особенностей:

- для лиц с нарушениями зрения: в печатной форме увеличенным шрифтом; в форме электронного документа; индивидуальные консультации с привлечением тифлосурдопереводчика; индивидуальные задания, консультации и др.

- для лиц с нарушениями слуха: в печатной форме; в форме электронного документа; индивидуальные консультации с привлечением сурдопереводчика; индивидуальные задания, консультации и др.

- для лиц с нарушениями опорно-двигательного аппарата: в печатной форме; в форме электронного документа; индивидуальные задания, консультации и др.

6 Фонд оценочных средств для проведения текущего контроля и промежуточной аттестации обучающихся по дисциплине (модулю)

В соответствии с требованиями ФГОС ВО для аттестации обучающихся на соответствие их персональных достижений планируемым результатам обучения по дисциплине (модулю) созданы фонды оценочных средств. Типовые контрольные задания, методические материалы, определяющие процедуры оценивания знаний, умений и навыков, а также критерии и показатели, необходимые для оценки знаний, умений, навыков и характеризующие этапы формирования компетенций в процессе освоения образовательной программы, представлены в Приложении 1.

7 Учебно-методическое и информационное обеспечение дисциплины (модуля)

7.1 Основная литература

1. Барский, А. Б., Искусственный интеллект и интеллектуальные системы управления : монография / А. Б. Барский. — Москва : Русайнс, 2022. — 185 с. — ISBN 978-5-4365-8166-8. — URL: https://book.ru/book/943706 (дата обращения: 18.06.2025). — Текст : электронный.

2. Зараменских, Е. П. Интернет вещей. Исследования и область применения : монография / Е.П. Зараменских, И.Е. Артемьев. — Москва : ИНФРА-М, 2024. — 188 с. — (Научная мысль). — DOI 10.12737/13342. - ISBN 978-5-16-019914-6. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2144319 (Дата обращения -18.06.2025)

3. Парфенов, Ю. П. Постреляционные хранилища данных : учебное пособие для вузов / Ю. П. Парфенов ; под научной редакцией Н. В. Папуловской. — Москва : Издательство Юрайт, 2025. — 97 с. — (Высшее образование). — ISBN 978-5-534-21173-3. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/559502 (дата обращения: 18.06.2025).

7.2 Дополнительная литература

1. Григорьев, Л. Ю. Petroleum Engineering Handbook. Upstream : учебное пособие / Л. Ю. Григорьев. — Ухта: УГТУ, 2021. — 228 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/209576 (дата обращения: 17.06.2025). — Режим доступа: для авториз. пользователей.

2. Дадян, Э. Г. Данные: хранение и обработка : учебник / Э. Г. Дадян. — Москва : ИНФРА-М, 2021. — 205 с. — (Высшее образование: Бакалавриат). - ISBN 978-5-16-016447-2. - Текст : электронный. - URL: https://znanium.ru/catalog/product/1149101 (Дата обращения -18.06.2025)

3. Лысенкова, С. Н. «Распределенные базы данных». Основы языка SQL : учебное пособие / С. Н. Лысенкова. — Брянск : Брянский ГАУ, 2022. — 36 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL:

https://e.lanbook.com/book/305006 (дата обращения: 17.06.2025). — Режим доступа: для авториз. пользователей.

4. Пальмов, С. В. Основы сбора и обработки больших данных : учебное пособие / С. В. Пальмов. — Самара : ПГУТИ, 2023. — 285 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/411830 (дата обращения: 17.06.2025). — Режим доступа: для авториз. пользователей.

7.3 Ресурсы информационно-телекоммуникационной сети "Интернет", включая профессиональные базы данных и информационно-справочные системы (при необходимости):

1. Образовательная платформа "ЮРАЙТ"
2. Электронно-библиотечная система "BOOK.ru"
3. Электронно-библиотечная система "ZNANIUM.COM"
4. Электронно-библиотечная система "ЛАНЬ"
5. Open Academic Journals Index (OAJI). Профессиональная база данных - Режим доступа: http://oaji.net/
6. Президентская библиотека им. Б.Н.Ельцина (база данных различных профессиональных областей) - Режим доступа: https://www.prlib.ru/
7. Информационно-справочная система "Консультант Плюс" - Режим доступа: http://www.consultant.ru/

8 Материально-техническое обеспечение дисциплины (модуля) и перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения

Основное оборудование:

- Коммутатор SuperStack 3 (16*10/100 19")
- Мультимедийный комплект №2 в составе:проектор Casio XJ-M146,экран 180*180,крепление потолочное
- Мультимедийный проектор Casio XJ-V2
- Облачный монитор 23" LG CAV42K
- Облачный монитор LG Electronics черный +клавиатура+мышь
- Сетевой монитор:Нулевой клиент Samsung SyncMaster NC240
- Уст-во бесп.питания UPS-3000

Программное обеспечение:

- Microsoft Office Professional Plus 2013 Russian
- Python
- Visual Studio

МИНОБРНАУКИ РОССИИ

ВЛАДИВОСТОКСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

НАУЧНО-ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР "ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ"

Фонд оценочных средств
для проведения текущего контроля
и промежуточной аттестации по дисциплине (модулю)

ОСНОВЫ DATA ENGINEERING

Направление и направленность (профиль)
09.04.03 Прикладная информатика. Искусственный интеллект и машинное обучение в управлении и принятии решений

Год набора на ОПОП
2024

Форма обучения
очная

Владивосток 2025

1 Перечень формируемых компетенций

Название ОПОП ВО, сокращенное	Код и формулировка компетенции	Код и формулировка индикатора достижения компетенции
09.04.03 «Прикладная информатика» (М-ПИ)		

Компетенция считается сформированной на данном этапе в случае, если полученные результаты обучения по дисциплине оценены положительно (диапазон критериев оценивания результатов обучения «зачтено», «удовлетворительно», «хорошо», «отлично»). В случае отсутствия положительной оценки компетенция на данном этапе считается несформированной.

2 Показатели оценивания планируемых результатов обучения

Таблица заполняется в соответствии с разделом 1 Рабочей программы дисциплины (модуля).

3 Перечень оценочных средств

Таблица 3 – Перечень оценочных средств по дисциплине (модулю)

Контролируемые планируемые результаты обучения		Контролируемые темы дисциплины	Наименование оценочного средства и представление его в ФОС	
			Текущий контроль	Промежуточная аттестация
Очная форма обучения				
РД1	Знание : Базовые технологии больших данных	1.1. Генерация данных	Собеседование	Проект
РД2	Умение : Выбор подходящих технологий	1.2. Инфраструктура больших данных	Собеседование	Проект
РД3	Знание : Архитектуры инфраструктуры хранения и обработки больших данных	1.3. MapReduce	Собеседование	Проект
РД4	Умение : Построение инфраструктуры больших данных	1.2. Инфраструктура больших данных	Собеседование	Проект
		1.3. MapReduce	Собеседование	Проект
РД5	Навык : Базовая настройка компонентов инфраструктуры больших данных	1.2. Инфраструктура больших данных	Собеседование	Проект
		1.4. Технологии Big Data	Собеседование	Проект

4 Описание процедуры оценивания

Качество сформированности компетенций на данном этапе оценивается по результатам текущих и промежуточных аттестаций при помощи количественной оценки, выраженной в баллах. Максимальная сумма баллов по дисциплине (модулю) равна 100 баллам.

Вид учебной деятельности	Оценочное средство		
	Вопросы для собеседования	Проект	Итого
Промежуточная аттестация		40	40
Практические занятия	50		50
Самостоятельная работа		10	10
Итого	50	40	100

Сумма баллов, набранных студентом по всем видам учебной деятельности в рамках дисциплины, переводится в оценку в соответствии с таблицей.

Сумма баллов по дисциплине	Оценка по промежуточной аттестации	Характеристика качества сформированности компетенции
от 91 до 100	«зачтено» / «отлично»	Студент демонстрирует сформированность дисциплинарных компетенций, обнаруживает всестороннее, систематическое и глубокое знание учебного материала, усвоил основную литературу и знаком с дополнительной литературой, рекомендованной программой, умеет свободно выполнять практические задания, предусмотренные программой, свободно оперирует приобретенными знаниями, умениями, применяет их в ситуациях повышенной сложности.
от 76 до 90	«зачтено» / «хорошо»	Студент демонстрирует сформированность дисциплинарных компетенций: основные знания, умения освоены, но допускаются незначительные ошибки, неточности, затруднения при аналитических операциях, переносе знаний и умений на новые, нестандартные ситуации.
от 61 до 75	«зачтено» / «удовлетворительно»	Студент демонстрирует сформированность дисциплинарных компетенций: в ходе контрольных мероприятий допускаются значительные ошибки, проявляется отсутствие отдельных знаний, умений, навыков по некоторым дисциплинарным компетенциям, студент испытывает значительные затруднения при оперировании знаниями и умениями при их переносе на новые ситуации.
от 41 до 60	«не зачтено» / «неудовлетворительно»	У студента не сформированы дисциплинарные компетенции, проявляется недостаточность знаний, умений, навыков.
от 0 до 40	«не зачтено» / «неудовлетворительно»	Дисциплинарные компетенции не сформированы. Проявляется полное или практически полное отсутствие знаний, умений, навыков.

5 Примерные оценочные средства

5.1 Примерный перечень вопросов по темам и для проведения собеседования

Перечень примерных вопросов на собеседование

Базовые технические вопросы

1. **В чем разница между ETL и ELT процессами?**
2. **Что такое data modeling и какие типы моделей данных вы знаете?** Расскажите о звездообразной схеме, схеме снежинки и их применении.
3. **Объясните концепцию нормализации данных.** Нормальные формы вы знаете, когда денормализация может быть полезна?
4. **Что такое data lake, data warehouse и data mart?** В чем их ключевые различия и применения?
5. **Как обеспечить качество данных в пайплайне?**
6. **Объясните разницу между OLTP и OLAP системами.**

7. **Что такое slowly changing dimensions (SCD)?** Опишите различные типы SCD и сценарии их применения.
8. **Какие стратегии партиционирования данных вы знаете?**
9. **Что такое идемпотентность в контексте пайплайнов данных?** Как обеспечивается идемпотентность процессов?
10. **Объясните концепцию data lineage.** Почему это важно и какие инструменты используются для его отслеживания?

Big Data и облачные технологии:

1. **Принципы работы MapReduce.** Каковы его ограничения и почему появились альтернативы?
2. **Hadoop экосистема.** Какие компоненты используются и для каких задач?
3. **В чем разница между batch и stream processing?** Выбрали между Spark Streaming и Kafka Streams?
4. **Объясните концепцию data partitioning в Spark.** Как правильно выбрать ключ партиционирования?
5. **Что такое Spark RDD, DataFrames и Datasets?**
6. **Как бороться с data skew в распределенных системах?** Приведите примеры решений.
7. **Объясните концепцию serverless в контексте обработки данных.** Каковы преимущества и недостатки AWS Lambda или Google Cloud Functions?
8. **Как реализовать CI/CD для дата-пайплайнов в облачной среде?** .
9. **Что такое data mesh и data fabric?** Как эти архитектуры меняют подход к работе с данными?
10. **Объясните преимущества и недостатки использования managed services вроде AWS Glue или Google BigQuery.**
11. **Как обеспечить безопасность данных в облачной инфраструктуре?**
12. **Что такое data governance и почему это важно в облачных средах?** Какие инструменты вы бы предпочли использовать?

Краткие методические указания

После выполнения каждой практической работы студент предоставляет **результаты выполнении задания и проходит собеседование по теме.**

Шкала оценки

№	Баллы	Описание
5	41-50	Студент демонстрирует умения на итоговом уровне: умеет свободно выполнять задания, предус мотренные программой, свободно оперирует приобретенными умениями, применяет их в ситуа циях повышенной сложности.
4	28-40	Студент демонстрирует умения на среднем уровне: освоил основные умения, но допускаются н езначительные ошибки, неточности, затруднения при аналитических операциях, переносе умен ий на новые, нестандартные ситуации.
3	19-27	Студент демонстрирует умения и навыки на базовом уровне: допускаются значительные ошибк и, проявляется отсутствие отдельных умений, навыков по дисциплинарной компетенции, испыт ываются значительные затруднения при оперировании умениями и при их переносе на новые с итуации.
2	10-18	Студент демонстрирует умения и навыки на уровне ниже базового: проявляется недостаточност ь умений и навыков.
1	0-9	Студентом проявляется полное или практически полное отсутствие умений и навыков, но прис утствует на занятии и пытается выполнить задание.

5.2 Проект

Проектное задание выполняется студентами индивидуальлно или малыми группами (до 3 человек). Проектное задание должно включать следующие разделы:

1. Источники данных, особенности располагаемых массивов данных, особенности работы с больших данными
2. Инфраструктура больших данных.
3. Технологии хранения располагаемых массивов данных
4. Технологии извлечения и обработки располагаемых массивов данных
5. Возможность и целесообразность применения технологий распределенных вычисления MapReduce,технлогиии распределенного хранение к располагаемому массиву данных

Задача 1. Установка сервера

Провести предварительную настройку операционной системы. Установить Java версии 6x или старше. Создать отдельной учетной записи для запуска Hadoop. Настроить доступ по ssh для управления узлами кластера hadoop.Установить Apache Hadoop.

Задача 2. Настройка Apache Hadoop

Задача3. Интеграция и обработка на платформе Hadoop

Краткие методические указания

Итоговая аттестация проводится по результатам выполнения проектного задания, которое включает все практические работы, которые объединены в единый проект , с учетом особенностей располагаемого (полученного студентом из открытых источников) массива данных. На итоговом занятии, студент защищает проектное задание, отвечает на вопросы, аргументирует решения

Шкала оценки

№	Баллы	Описание
5	35-40	Студент ответил на заданные вопросы. Студент демонстрирует знания на высоком уровне. Дем онстрирует способность самостоятельно реализовать полный цикл создания, хранения и обраб отки массивов больших данных
4	29-34	Студент демонстрирует знания на среднем уровне. Демонстрирует способность реализовать по лный цикл создания, хранения и обработки массивов больших данных
3	22-28	Студент демонстрирует знания на базовом уровне. Способен реализовать частично цикл создан ия, хранения и обработки массивов больших данных
2	10-21	Студент демонстрирует знания на уровне ниже базового.
1	0-9	Студентом проявляется полное или практически полное отсутствие знаний, но присутствовал н а занятии и пытался ответить на вопросы.

Ключи ФОС Data engeniring

Базовые технические вопросы

1. В чем разница между ETL и ELT процессами?

ETL (Extract, Transform, Load):

Данные сначала извлекаются (Extract) из источников, затем трансформируются (Transform) в промежуточном слое (например, в ETL-инструменте), и только после этого загружаются (Load) в целевую систему (например, Data Warehouse).

Подходит для структурных данных и случаев, когда трансформации сложные и требуют значительных ресурсов.

ELT (Extract, Load, Transform):

Данные извлекаются (Extract), сразу загружаются (Load) в целевую систему (например, облачное хранилище или DWH), а трансформации (Transform) выполняются уже внутри неё.

Подходит для больших объёмов данных (Big Data), облачных хранилищ и случаев, когда сырые данные нужны для анализа.

Ключевое отличие:

ETL трансформирует данные до загрузки, а ELT — после. ELT более гибкий и масштабируемый, особенно в облачных средах.

2. Что такое data modeling и какие типы моделей данных вы знаете?

Data modeling — процесс создания структуры данных для эффективного хранения и анализа.

Основные типы моделей:

Звездообразная схема (Star Schema):

Центральная таблица фактов, связанная с таблицами измерений (dimensions).

Простота и высокая производительность для запросов.

Пример: продажи (факты) + товары, клиенты, время (измерения).

Снежинка (Snowflake Schema):

Усложнённая звезда, где измерения нормализованы (разбиты на подтаблицы).

Экономит место, но усложняет запросы.

Пример: измерение "продукт" разбито на "категория" и "подкатегория".

Реляционная модель:

Таблицы с связями (1:1, 1:N, N:M), нормализованные для минимизации дублирования.

Применение:

Звезда — для аналитики (OLAP), снежинка — для сложных структур с экономией места.

3. Объясните концепцию нормализации данных.

Нормализация — устранение дублирования и аномалий через разбиение таблиц на связанные сущности.

Основные нормальные формы:
1. 1NF: Все атрибуты атомарны (неделимы), нет повторяющихся групп.
2. 2NF: Выполнена 1NF + нет частичных зависимостей от составного ключа.
3. 3NF: Выполнена 2NF + нет транзитивных зависимостей (атрибуты зависят только от первичного ключа).

Денормализация:
Намеренное объединение таблиц для ускорения запросов (например, в аналитических системах OLAP).

4. Что такое data lake, data warehouse и data mart?

Data Lake:
Хранилище сырых данных (структурированных, полуструктурированных, неструктурированных).
Гибкость, но требует обработки перед анализом (например, Hadoop, S3).

Data Warehouse (DWH):
Структурированное хранилище для аналитики (ETL/ELT, схемы "звезда"/"снежинка").
Оптимизирован для SQL-запросов (например, Redshift, BigQuery).

Data Mart:
Подмножество DWH для конкретного отдела (например, финансы или маркетинг).

Различия:

Характеристика	Data Lake	Data Warehouse	Data Mart
Данные	Сырые	Очищенные	Тематические
Структура	Гибкая	Жёсткая	Узкая

5. Как обеспечить качество данных в пайплайне?

Валидация: Проверка на корректность (форматы, диапазоны).
Очистка: Удаление дубликатов, заполнение пропусков.
Мониторинг: Алёрты при аномалиях (например, резком падении объёма данных).
Тестирование: Юнит-тесты для трансформаций.
Документация: Описание метаданных и lineage.

Инструменты: Great Expectations, dbt, Apache Griffin.

6. Разница между OLTP и OLAP системами.

OLTP (Online Transaction Processing):
Операционные системы для быстрых транзакций (например, банковские операции).
Примеры: MySQL, PostgreSQL.

OLAP (Online Analytical Processing):
Аналитические системы для сложных запросов по историческим данным.
Примеры: ClickHouse, Snowflake.

Сравнение:

Характеристика	OLTP	OLAP
Назначение	Транзакции	Аналитика
Данные	Текущие	Исторические
Оптимизация	Запись	Чтение

7. Что такое slowly changing dimensions (SCD)?
SCD — методы обработки изменений в измерениях (например, смена адреса клиента).

Типы SCD:
Тип 1: Перезапись старых данных (история не сохраняется).
Тип 2: Добавление новой строки с метками времени (актуальности).
Тип 3: Хранение ограниченной истории (например, предыдущее значение в отдельном столбце).

Применение:
Тип 2 — для аудита и анализа истории.

8. Стратегии партиционирования данных.
По диапазону: Разделение по значениям (например, дата: `2023-01`, `2023-02`).
По списку: Группировка по категориям (например, регион: EU, US).
По хэшу: Равномерное распределение для балансировки нагрузки.
По времени: Для временных рядов (логи, события).

Цель: Ускорение запросов и управление большими таблицами.

9. Идемпотентность в пайплайнах данных.
Идемпотентность — свойство операции, при котором повторное выполнение не меняет результат (например, `UPDATE` с условием).

Как обеспечить:
Использование уникальных ключей для вставки.
Проверка на существование данных перед обработкой.
Транзакции и механизмы "upsert" (INSERT ... ON CONFLICT UPDATE).

10. Концепция data lineage.
Data lineage — отслеживание происхождения данных, их перемещения и трансформаций в пайплайне.

Важность:
Аудит и соответствие регуляциям (GDPR).
Поиск причин ошибок.
Оптимизация процессов.
Инструменты: Apache Atlas, Collibra, Alation.

Big Data и облачные технологии

1. Принципы работы MapReduce. Каковы его ограничения и почему появились альтернативы?

Принципы:
- Map: Разделение данных на части и параллельная обработка (например, фильтрация или преобразование).
- Reduce: Агрегация результатов из этапа Map (например, суммирование или группировка).

Ограничения:
- Высокая задержка: Запись промежуточных данных на диск замедляет процесс.
- Сложность: Написание кода для сложных пайплайнов требует много усилий.
- Не подходит для стриминга: Только пакетная обработка (batch).

Альтернативы:
- Apache Spark: Использует оперативную память (in-memory), поддерживает стриминг и SQL.
- Flink: Оптимизирован для потоковой обработки с низкой задержкой.

2. Hadoop экосистема. Какие компоненты используются и для каких задач?

Основные компоненты:
- HDFS (Hadoop Distributed File System): Распределённое хранение данных.
- YARN: Управление ресурсами кластера.
- MapReduce: Пакетная обработка (устаревший, заменён на Spark).
- Hive: SQL-интерфейс для запросов к данным в HDFS.
- HBase: NoSQL-база для реального доступа к данным.
- Spark: Быстрая обработка in-memory.
- Kafka: Стриминг данных.

Пример использования:
- HDFS + Spark для ETL, Kafka для потоковых данных, Hive для аналитики.

3. Разница между batch и stream processing. Выбор между Spark Streaming и Kafka Streams?

- Batch: Обработка данных порциями (например, раз в день). Подходит для сложных аналитических задач.
- Stream: Обработка в реальном времени (например, мониторинг транзакций).

Spark Streaming vs. Kafka Streams:

Критерий	Spark Streaming	Kafka Streams
Масштабируемость	Подходит для больших кластеров	Лучше для небольших реальных задач
Задержка	Минимум 1 секунда	Миллисекунды
Интеграция	Работает с HDFS, Kafka	Только Kafka

Выбор:
- Если уже используете Kafka и нужна низкая задержка — Kafka Streams.
- Для сложной аналитики и интеграции с другими системами — Spark Streaming.

4. Концепция data partitioning в Spark. Как выбрать ключ партиционирования?

Partitioning — разделение данных на части для параллельной обработки.

Как выбрать ключ:
- Равномерное распределение: Избегайте skew (например, хэш от ID вместо категории с малым числом значений).
- Частые фильтры: Партиционируйте по полям, которые часто используются в WHERE.
- JOIN-операции: Одинаковые ключи в соединяемых таблицах ускоряют JOIN.

Пример:
```python
df.repartition(100, "date")  # 100 партиций по дате
```

5. Spark RDD, DataFrames и Datasets

- RDD (Resilient Distributed Dataset):
  - Низкоуровневая абстракция, неизменяемая и распределённая.
  - Подходит для ручной оптимизации.
- DataFrame:
  - Табличная абстракция с оптимизацией (Spark SQL).
  - Поддержка SQL и встроенных функций (например, `groupBy`).
- Dataset:
  - Типизированная версия DataFrame (только для Scala/Java).

Когда использовать:
- RDD — для сложных низкоуровневых операций.
- DataFrame/Dataset — для аналитики и SQL.

6. Как бороться с data skew в распределённых системах?

Data skew — неравномерное распределение данных (например, 80% записей в одной партиции).

Решения:

- Солянка (Salting): Добавление случайного префикса к ключу (`key -> salted_key`).
- Увеличение партиций: `repartition(1000)`.
- Локальные JOIN: `broadcast` для маленьких таблиц.

Пример:
```python
from pyspark.sql.functions import concat, lit, rand
df = df.withColumn("salted_key", concat("key", lit("_"), (rand() * 10).cast("int")))
```

7. Serverless в обработке данных. AWS Lambda vs. Google Cloud Functions

Serverless — выполнение кода без управления серверами (автомасштабирование, оплата за использование).

Сравнение:

Критерий	AWS Lambda	Google Cloud Functions
Макс. время	15 минут	9 минут
Интеграция	S3, DynamoDB	BigQuery, Pub/Sub
Языки	Python, Node.js, Java	Python, Node.js, Go

Плюсы serverless:
- Нет инфраструктурных затрат.
- Автомасштабирование.

Минусы:
- Холодный старт (задержка при первом вызове).
- Ограничения по времени выполнения.

8. CI/CD для дата-пайплайнов в облаке

Подход:
- Тестирование: Юнит-тесты (например, pytest), интеграционные тесты с данными.
- Развёртывание: Terraform для инфраструктуры, Airflow для оркестрации.
- Инструменты:
  - GitHub Actions/GitLab CI для запуска тестов.
  - Docker для контейнеризации.
  - Jenkins для сложных пайплайнов.

Пример:
```yaml
# GitHub Actions для дата-пайплайна
jobs:
  test:
    runs-on: ubuntu-latest
    steps:
      - run: pytest tests/
```

```
  deploy:
    needs: test
    run: terraform apply
```

9. Data Mesh и Data Fabric
- Data Mesh:
  - Декомпозиция данных по доменам (например, финансы, маркетинг).
  - Каждая команда отвечает за свой домен.
  - Инструменты: Kafka, Domain-Driven Design.
- Data Fabric:
  - Единый слой для доступа к данным через метаданные и AI.
  - Пример: Informatica, IBM Cloud Pak.

Разница:
Data Mesh — организационная философия, Data Fabric — технологическая реализация.

10. Managed services: AWS Glue vs. Google BigQuery

AWS Glue:
- Плюсы: Интеграция с S3, Spark под капотом.
- Минусы: Сложность настройки, высокая стоимость при больших объёмах.

Google BigQuery:
- Плюсы: Мгновенное масштабирование, SQL-интерфейс.
- Минусы: Vendor lock-in, стоимость при частых запросах.

Выбор:
- Для сложного ETL — Glue.
- Для аналитики — BigQuery.

11. Безопасность данных в облаке
- Шифрование: AES-256 для данных в rest и TLS для передачи.
- IAM: Ролевой доступ (например, AWS IAM, GCP IAM).
- Мониторинг: AWS CloudTrail, Google Security Command Center.
- GDPR/Compliance: Маскирование PII-данных (например, с помощью Apache Ranger).

12. Data Governance в облаке

Data Governance — управление качеством, доступом и безопасностью данных.

Инструменты:
- Open Source: Apache Atlas, Marquez.
- Проприетарные: Collibra, Alation.

Почему важно:
- Соответствие регуляциям (GDPR, HIPAA).
- Предотвращение утечек и ошибок.

Ключи (критерии) для проектного задания)

Критерий	**Описание**
Анализ качества набора больших данных	Выбор и обоснование выбора методов анализа качества и характера данных располагаемого набора данных
Разметка данных в соответствии с классом задачи, моделью данных	Разметка данных соответствующая задаче, целям построения интеллектуальной модели
Инфраструктура больших данных	Аргументированные выводы о технологии извлечения и обработки располагаемых массивов данных
Технологии хранения располагаемых массивов данных	Выбор технологии хранения располагаемых массивов данных
Использование инструментария	Установка сервера: предварительную настройку операционной системы; установка Java версии 6х или старше; создатние отдельной учетной записи; настройка доступа по ssh для управления узлами кластера hadoop; установка Apache Hadoop
Достигнутые результаты	Умение студента реализовать полный цикл создания, хранения и обработки массивов больших данных средствами свободной программной платформы Hadoop