Оценочные материалы для промежуточной аттестации по дисциплине

Название дисциплины «Большие данные»

Код, направление подготовки	09.03.02 Информационные системы и технологии
Направленность (профиль)	Безопасность информационных систем и технологий
Форма обучения	Очная
Кафедра-разработчик	Информатики и вычислительной техники
Выпускающая кафедра	Информатики и вычислительной техники

Вопросы для зачёта

1. Дайте определение понятия «Большие данные» Перечислите и кратко охарактеризуйте основные характеристики больших данных по модели 5V

2. В чём заключаются принципиальные различия между традиционными реляционными СУБД и технологиями обработки больших данных? Приведите примеры задач, которые эффективнее решаются с помощью Big Data.

3. Опишите архитектуру распределённой файловой системы HDFS в составе Apache Hadoop. Какие проблемы она решает и как обеспечивается отказоустойчивость?

4. Сравните фреймворки Apache Hadoop MapReduce и Apache Spark. Укажите ключевые преимущества Spark по производительности, удобству разработки и областям применения.

5. Перечислите основные типы NoSQL-баз данных (ключ-значение, документные, колоночные, графовые). Для каждого типа приведите пример СУБД и укажите характерные особенности и сценарии использования.

6. Что такое Data Lake, Data Warehouse и Data Lakehouse? Сравните эти концепции по ключевым характеристикам (схема, стоимость хранения, поддержка ACID, инструменты анализа) и укажите, когда целесообразно применять каждую.

7. Объясните различия между пакетной (batch) и потоковой (stream) обработкой данных. Приведите примеры задач и инструментов для каждого подхода.

8. Опишите Lambda-архитектуру и Kappa-архитектуру. В чём их сходства и различия? В каких сценариях предпочтительна каждая из них?

9. Какую роль играет Apache Kafka в экосистеме Big Data? Опишите основные абстракции (топик, партиция, продюсер, консьюмер, брокер) и принципы работы.

10. В чём различие между процессами ETL и ELT? Приведите примеры сценариев, когда предпочтительнее использовать каждый из подходов.

11. Опишите основные возможности Apache Spark для обработки больших данных. Что представляют собой абстракции RDD, DataFrame и Dataset? В каких случаях какую из них целесообразно использовать?

12. Как технологии Big Data применяются для решения задач машинного обучения? Назовите популярные инструменты и библиотеки (Spark MLlib, TensorFlow/PyTorch на Spark и др.).

13. Какие основные вызовы и проблемы в области **безопасности**, **конфиденциальности и этики данных возникают при работе с большими данными? Приведите примеры подходов к их решению.

14. Охарактеризуйте возможности облачных платформ для работы с Big Data на примере AWS, Microsoft Azure или Google Cloud Platform. Перечислите ключевые сервисы каждой платформы.

15. Что включает в себя Data Governance (управление данными) в проектах Big Data? Почему управление данными особенно важно при работе с большими объёмами данных?

Типовые задания для контрольной работы:

Теоретические вопросы

1. Перечислите и кратко поясните все пять характеристик модели 5V больших данных. Приведите пример для каждой характеристики.

2. В чём принципиальное отличие Data Lake от Data Warehouse? В каких случаях предпочтительнее использовать каждый из них?

3. Объясните назначение и принцип работы компонентов HDFS (NameNode, DataNode, Secondary NameNode). Что произойдёт при выходе из строя NameNode?

4. Сравните модели обработки данных MapReduce и Apache Spark. Укажите минимум три ключевых преимущества Spark.

Практические задания

Дан датасет
Напишите код, который:

а) Загружает данные в Spark DataFrame.
б) Выполняет очистку (удаление дубликатов, обработка NULL, фильтрация некорректных дат).
в) Агрегирует данные:
- Сумма и количество транзакций по категориям за каждый день.
- Топ-10 пользователей по какому-либо критерию.
г) Строит простую модель линейной регрессии для предсказания суммы следующей транзакции пользователя (по его истории)