Тестовое задание для диагностического тестирования по дисциплине:

Информатика

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

1 семестр

Проверяемая компетенция	№	Задание	Варианты ответов
	1	Какие пять основных характеристик определяют понятие «большие данные»?	a) Volume, Velocity, Variety, Viscosity, Volatility b) Volume, Velocity, Variety, Veracity, Value c) Volume, Velocity, Validity, Veracity, Visibility d) Volume, Variety, Velocity, Validity, Vector
	2	Что такое Data Lake?	a) Хранилище только структурированных и очищенных данных b) Репозиторий сырых данных в исходных форматах (схема-on-read) c) Реляционная база данных с жёсткой схемой d) Система только для машинного обучения
	3	Каково основное назначение HDFS в экосистеме Hadoop?	a) Выполнение машинного обучения b) Распределённое файловое хранилище с репликацией c) Потоковая обработка событий в реальном времени d) Визуализация данных
	4	Что обозначает аббревиатура RDD в Apache Spark?	a) Relational Data Definition b) Resilient Distributed Dataset c) Real-time Data Dashboard d) Remote Distributed Database
	5	Какой компонент Apache Spark предназначен для выполнения SQL-подобных запросов?	a) Spark Streaming b) Spark SQL c) MLlib

			d) GraphX
	6	Какой инструмент лучше всего подходит для обработки потоковых данных в реальном времени?	a) Hadoop MapReduce b) Spark Structured Streaming c) HDFS d) HBase
	7	Что такое ETL-процесс?	a) Extract, Transform, Load b) Execute, Test, Load c) Extract, Transfer, Link d) Encode, Transform, Load
	8	Какой алгоритм в Spark MLlib чаще всего используется для кластеризации?	a) Linear Regression b) K-Means c) Decision Tree d) Logistic Regression
	9	Может ли Apache Spark работать без Hadoop HDFS?	a) Нет, Spark работает только поверх HDFS b) Да, Spark поддерживает множество хранилищ (S3, GCS, локальная FS и др.) c) Только с Cassandra d) Только в облаке Yandex
	10	Что такое partitioning в Spark DataFrame?	a) Удаление части данных b) Разбиение данных на логические части для параллельной обработки c) Сжатие данных d) Шифрование данных
	11	Какое ключевое преимущество DataFrame над RDD?	a) Более низкоуровневый API b) Автоматическая оптимизация запросов через Catalyst Optimizer c) Отсутствие схемы данных d) Более медленная обработка
	12	Какой формат файлов рекомендуется для хранения аналитических данных в Spark?	a) CSV b) Parquet c) TXT d) JSON
	13	Что означает термин fault-tolerance в Spark?	a) Устойчивость к сбоям за счёт репликации b) Увеличение скорости обработки c) Сжатие данных d) Визуализация результатов
	14	Какая база данных относится к колоночным NoSQL-системам?	a) MongoDB b) HBase c) Redis d) Neo4j
	15	Что такое Data Skew в Spark?	a) Равномерное распределение данных b) Неравномерное распределение данных по частям c) Сжатие данных

			d) Удаление дубликатов
	16	Какой процесс обычно следует за Extract в традиционном ETL?	a) Load b) Transform c) Analyze d) Visualize
	17	Что важно учитывать при работе с большими данными в РФ?	a) Только технические характеристики b) ФЗ-152 о персональных данных c) Только объём данных d) Только скорость обработки
	18	Для каких данных лучше использовать Data Lake?	a) Только очищенных и структурированных b) Сырых и разнородных данных в больших объёмах c) Только транзакционных данных d) Только для реляционных таблиц
	19	Какой компонент Hadoop отвечает за распределение ресурсов?	a) HDFS b) YARN c) MapReduce d) Hive
	20	Что лучше всего подходит для сбора и передачи событий в реальном времени?	a) HDFS b) Apache Kafka c) Spark SQL d) Parquet