Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Новосибирский государственный технический университет»
Кафедра автоматизированных систем управления

Паспорт экзамена

по дисциплине «Технологии и решения для анализа больших данных», 3 семестр

1. Методика оценки

Экзамен проводится в письменной форме, по тестам. В ходе экзамена преподаватель вправе задавать студенту дополнительные вопросы из общего перечня (п. 4).

Пример теста **для** экзамена

Вопрос № 1
Какие СУБД полностью полагаются на оперативную память при хранении информации:
а) Oracle Exalytics
б) SAP HANA
в) BigTable
г) HBase

Вопрос № 2
Для чего аналитику необходима «песочница»?
а) Для высокопроизводительной аналитики за счёт использования оперативной памяти и inDB операций.
б) Для хранения всех полученных от заказчика данных.
в) Для построения отчётов о результатах анализа
г) Для снижения затрат, связанных с репликацией данных

2. Критерии оценки

- Ответ на тест считается **неудовлетворительным**, если студент при ответе на вопросы не дает определений основных понятий, не способен показать причинно-следственные связи явлений, при решении задачи допускает принципиальные ошибки, оценка составляет *0 баллов.*
- Ответ на тест засчитывается на **пороговом** уровне, если студент при ответе на вопросы дает определение основных понятий, может показать причинно-следственные связи явлений, при решении задачи допускает непринципиальные ошибки, например, вычислительные,
  оценка составляет *10 баллов.*
- Ответ на тест засчитывается на **базовом** уровне, если студент при ответе на вопросы формулирует основные понятия, законы, дает характеристику процессов, явлений, проводит анализ причин, условий, может представить качественные характеристики процессов, не допускает ошибок при решении задачи, оценка составляет *20 баллов.*
- Ответ на тест засчитывается на **продвинутом** уровне, если студент при ответе на вопросы проводит сравнительный анализ подходов, проводит комплексный анализ, выявляет проблемы, предлагает механизмы решения, способен представить количественные характеристики определенных процессов, приводит конкретные примеры из практики, не допускает ошибок и способен обосновать выбор метода решения задачи, оценка составляет 40 *баллов.*

3. Шкала оценки

Экзамен считается сданным, если сумма баллов по всем заданиям билета оставляет не менее 10 баллов (из 40 возможных).

В общей оценке по дисциплине экзаменационные баллы учитываются в соответствии с правилами балльно-рейтинговой системы, приведенными в рабочей программе дисциплины, следующим образом: общая оценка по дисциплине складывается из баллов, полученных на экзамене (максимум 40), и баллов за выполнение практической части (максимум 30) и РГЗ (максимум 30).

Таблица соответствия баллов, традиционной оценки и буквенной оценки ECTS приведена ниже.

Диапазон баллов рейтинга	Оценка ECTS	Традиционная шкала оценки	
96 -100	A	отлично	зачтено
93 - 96	A		
90 - 93	A-		
86 - 89	B		
83 - 86	B	хорошо	
80 - 83	B-		
76 - 79	C		
73 - 76	C		
70 - 73	C-		
66 - 69	D	удовлетворительно	зачтено
63 - 66	D		
60 - 63	D-		
50 - 59	E		
25 - 49	FX	неудовлетворительно	не зачтено
0 - 24	F		

4. **Вопросы теста к** экзамену **по дисциплине «**Технологии и решения для анализа больших данных**»**

Вариант 1

1 Hadoop это:
а) Набор утилит, и программный каркас для выполнения распределённых программ, работающих на кластерах.
б) Распределённая СУБД, позволяющая обрабатывать большие данные.
в) Язык выполнения заданий в парадигме MapReduce.
г) Распределённая файловая система, предназначенная для хранения файлов большого объёма

2 Начиная с каких размеров данных обоснованно применение кластера Hadoop для хранения данных?
а) 100Гб
б) 1Тб
в) 100Тб
г) 1Пб

3 Пример благоразумного использования Hadoop

а) Анализ 10 Гб данных.
б) Ежедневное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт).
в) Посекундное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт).
г) Построение графика пульса пациента в реальном времени.

4 Выберите верное утверждение:
а) Data Warehouse создаются для проверки гипотез при анализе больших данных.
б) «Песочница» используется для снижения нагрузки на основной Data Warehouse.
в) Каждый Data Warehouse должен содержать «песочницу».
г) «Песочница» необходима для любого процесса аналитики.

5 Какие из следующих средств разумно использовать для анализа данных, представленных единственным csv - файлом размера более 100Гб:
а) Hadoop
б) Data Warehouse
в) «Песочница»
г) Python

6 Для чего аналитику необходима «песочница»?
а) Для высокопроизводительной аналитики за счёт использования оперативной памяти и inDB операций.
б) Для хранения всех полученных от заказчика данных.
в) Для построения отчётов о результатах анализа
г) Для снижения затрат, связанных с репликацией данных

7 В чём преимущество колоночно-ориентированных СУБД?
а) Они позволяют выполнять более сложные SQL-запросы по сравнению с реляционными СУБД
б) Они позволяют динамически дополнять содержание записей новыми полями
в) Они имеют более гибкие возможности аналитики.
г) Они позволяют эффективно делать межколоночные сравнения.

8 Какие СУБД полностью полагаются на оперативную память при хранении информации:
а) Oracle Exalytics
б) SAP HANA
в) BigTable
г) HBase

9 Какие из следующих технологий СУБД не используют принцип MapReduce
а) Hadoop
б) Cassandra
в) HDInsight
г) Redis

10 Выберите одно неверное высказывание про MapReduce:
а) Интерфейс для массово-параллельной обработки данных, где вычисления производятся на узлах, где информация изначально была сохранена
б) MapReduce –это две операции: распределения и сборки данных
в) MapReduce был придуман разработчиками Hadoop
г) MapReduce был анонсирован разработчиками Google

11 Принцип MapReduce состоит в том, чтобы
а) Производить вычисления на узлах, где информация изначально была сохранена
б) Использовать вычислительные мощности систем хранения
в) Использовать функциональное программирование для решения задач массивно-параллельной обработки
г) Использовать две операции - две операции: распределение и сборка данных

12 Отметьте неверное понимание Varietyв контексте характеристик Big Data
а) Высокая скорость генерирования данных.
б) Разные типы данных в колонках таблиц реляционных СУБД.
в) Разнообразие отраслей, являющихся источниками данных.
г) Разнообразие типов данных, включающих в себя структурированные, полуструктурированные и нест руктурированные.

13 Выберите неверное высказывание
а) Большие объёмы данных приводят к слабой их структуризации, поэтому появляется такое разнообразие данных.
б) Увеличившаяся производительность телекоммуникационных каналов привела к росту объёмов передаваемой информации.
в) Удешевление систем хранения на единицу информации привело к росту рынка больших данных.
г) Большое разнообразие источников данных

14 Перечислите четыре основных характеристики Big Data:
а) Virtualization, Volume, Variability, Vehicle
б) Variety, Velocity, Volume, Value
в) Verification, Volume, Velocity, Visualization
г) Video, Value, Variety , Volume

15 Отметьте те из вариантов, в которых данные структурированы:
а) Данные о продажах компании, представленные в виде помесячных отчётов в формате MS Word.
б) Таблица с ежедневными показаниями температуры помещения за год в файле формата csv.
в) Текст книги А.С. Макаренко Педагогическая поэма, представленный в формате PDF.
г) Библиотека фильмов, представленных в формате mpeg

Вариант 2

1 Выберите неверный ответ:
а) Большие данные – это данные объёма свыше 1 Тб
б) Проблема больших данных – это проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна.
в) Большие данные – это тренд в области ИТ, подогреваемый маркетинговыми кампаниями крупных вендоров.
г) Большие данные как правило не структурированы.

2 Выберите верный ответ
а) Большие данные – это обработка или хранение более 1 Тб информации.

б) Проблема больших данных – это такая проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна.
в) Большие данные – это огромная PR-акция крупных вендоров и не более того.
г) Большие данные – это явление, когда цифровые данные наиболее полно представляют изучаемый объект.

3 Отметьте значимые события, повлиявшие на формирование тренда больших данных:
а) Разработка Hadoop
б) Изобретениепринципа MapReduce
в) Разработка языка Pyhton
г) Победа Deepblue в матче с Г.Каспаровым.

4 Укажите фактор, способствовавший появлению тренда больших данных
а) Маркетинговые кампании крупных корпораций
б) Снижение издержек на хранение данных
в) Появление новых технологий обработки потоковых данных
г) Выпуск баз данных с обработкой данных в памяти

5 На каком из этапов процесса CRISP-DM происходит проверка гипотез?
а) Понимание бизнеса (Business understanding)
б) Понимание данных (Data Understanding)
в) Моделирование (Modeling)
г) Оценка (Evaluation)

6 Выберите верное утверждение:
а) Data Warehouse создаются для проверки гипотез при анализе больших данных.
б) «Песочница» используется для снижения нагрузки на основной Data Warehouse.
в) Каждый Data Warehouse должен содержать «песочницу».
г) «Песочница» необходима для любого процесса аналитики.

7 Отметьте неверное понимание Varietyв контексте характеристик Big Data
а) Высокая скорость генерирования данных.
б) Разные типы данных в колонках таблиц реляционных СУБД.
в) Разнообразие отраслей, являющихся источниками данных.
г) Разнообразие типов данных, включающих в себя структурированные, полуструктурированные и нест руктурированные.

8 Какие из следующих средств разумно использовать для анализа данных, представленных единственным csv - файлом размера более 100Гб:
а) Hadoop
б) Data Warehouse
в) «Песочница»
г) Python

9 Пример благоразумного использования Hadoop
а) Анализ 10 Гб данных.
б) Ежедневное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт).
в) Посекундное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт).
г) Построение графика пульса пациента в реальном времени.

10 Какие СУБД полностью полагаются на оперативную память при хранении информации:
а) Oracle Exalytics
б) SAP HANA
в) BigTable
г) HBase

11 Для чего аналитику необходима «песочница»?
а) Для высокопроизводительной аналитики за счёт использования оперативной памяти и inDB операций.
б) Для хранения всех полученных от заказчика данных.
в) Для построения отчётов о результатах анализа
г) Для снижения затрат, связанных с репликацией данных

12 Укажите фактор, способствовавший появлению тренда больших данных
а) Маркетинговые кампании крупных корпораций
б) Снижение издержек на хранение данных
в) Появление новых технологий обработки потоковых данных
г) Выпуск баз данных с обработкой данных в памяти

13 Hadoop это:
а) Набор утилит, и программный каркас для выполнения распределённых программ, работающих на кластерах.
б) Распределённая СУБД, позволяющая обрабатывать большие данные.
в) Язык выполнения заданий в парадигме MapReduce.
г) Распределённая файловая система, предназначенная для хранения файлов большого объёма

14 Выберите неверное высказывание
а) Большие объёмы данных приводят к слабой их структуризации, поэтому появляется такое разнообразие данных.
б) Увеличившаяся производительность телекоммуникационных каналов привела к росту объёмов передаваемой информации.
в) Удешевление систем хранения на единицу информации привело к росту рынка больших данных.
г) Большое разнообразие источников данных

15 Начиная с каких размеров данных обоснованно применение кластера Hadoop для хранения данных?
а) 100Гб
б) 1Тб
в) 100Тб
г) 1Пб

ТАБЛИЦА ПРАВИЛЬНЫХ ОТВЕТОВ
НА ТЕСТОВЫЕ ЗАДАНИЯ
ДЛЯ ПРОВЕРКИ ЗНАНИЙ СТУДЕНТОВ

Вариант 1

Номер задания	Вариант ответа				Номер задания	Вариант ответа			
	А	Б	В	Г		А	Б	В	Г
1	+				13	+			
2			+	+	14		+		
3			+		15		+		
4		+			16				
5	+			+	17				
6	+			+	18				
7		+			19				
8	+	+			20				
9				+	21				
10	+		+		22				
11	+	+		+	23				
12	+	+	+		24				

ТАБЛИЦА ПРАВИЛЬНЫХ ОТВЕТОВ
НА ТЕСТОВЫЕ ЗАДАНИЯ
ДЛЯ ПРОВЕРКИ ЗНАНИЙ СТУДЕНТОВ

Вариант 2

Номер задания	Вариант ответа				Номер задания	Вариант ответа			
	А	Б	В	Г		А	Б	В	Г
1	+				13	+			
2		+			14	+			
3	+	+			15			+	+
4	+	+			16				
5			+		17				
6		+			18				
7	+	+	+		19				
8	+		+	+	20				
9			+		21				
10	+	+			22				
11	+				23				
12					24				