Приложение к РПД

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

для проведения текущей и промежуточной аттестации

по учебной дисциплине (модулю)

«Технологии обработки больших данных»

для направления подготовки (уровень магистратуры):
09.04.01 «Информатика и вычислительная техника»

Направленность «Искусственный интеллект в автоматизированных системах обработки информации и управления»

Автор программы:

Абдеева Н.А.

Рецензент:

_____

Утверждена на заседании кафедры ПИМ «Прикладной информатики и математики»

Протокол №_10__ от «_26_» ___06____ 2022 г.

Заведующий кафедрой ПИМ

Глазырина И.П.. _____

Декан энергетического факультета

Батухтин А.Г. _____

Согласовано:

_____

Оглавление

1. ПЕРЕЧЕНЬ КОМПЕТЕНЦИЙ С УКАЗАНИЕМ ЭТАПОВ ИХ ФОРМИРОВАНИЯ В ПРОЦЕССЕ ОСВОЕНИЯ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ 4
2. ОПИСАНИЕ ПОКАЗАТЕЛЕЙ И КРИТЕРИЕВ ОЦЕНИВАНИЯ КОМПЕТЕНЦИЙ НА РАЗЛИЧНЫХ ЭТАПАХ ИХ ФОРМИРОВАНИЯ, ОПИСАНИЕ ШКАЛ ОЦЕНИВАНИЯ 5
3. ТИПОВЫЕ КОНТРОЛЬНЫЕ ЗАДАНИЯ ИЛИ ИНЫЕ МАТЕРИАЛЫ, НЕОБХОДИМЫЕ ДЛЯ ОЦЕНКИ ЗНАНИЙ, УМЕНИЙ, НАВЫКОВ И (ИЛИ) ОПЫТА ДЕЯТЕЛЬНОСТИ, ХАРАКТЕРИЗУЮЩИХ ЭТАПЫ ФОРМИРОВАНИЯ КОМПЕТЕНЦИЙ В ПРОЦЕССЕ ОСВОЕНИЯ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ 10
4. МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ, ОПРЕДЕЛЯЮЩИЕ ПРОЦЕДУРЫ ОЦЕНИВАНИЯ ЗНАНИЙ, УМЕНИЙ, НАВЫКОВ И (ИЛИ) ОПЫТА ДЕЯТЕЛЬНОСТИ, ХАРАКТЕРИЗУЮЩИХ ЭТАПЫ ФОРМИРОВАНИЯ КОМПЕТЕНЦИЙ. 11

4.1. Примеры (макеты) методических материалов, определяющих процедуры оценивания знаний, умений, навыков и (или) опыта деятельности. 12

4.2. Процедуры оценивания знаний, умений, навыков, формы и организация текущего контроля успеваемости и промежуточной аттестации обучающихся 30

1. ПЕРЕЧЕНЬ КОМПЕТЕНЦИЙ С УКАЗАНИЕМ ЭТАПОВ ИХ ФОРМИРОВАНИЯ В ПРОЦЕССЕ ОСВОЕНИЯ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Перечень компетенций (планируемых результатов освоения образовательной программы), выявленных в матрице компетенций, представлен в таблице 1 рабочей программы дисциплины совместно с планируемыми результатами обучения по дисциплине, а также в таблице 1 фонда оценочных средств (раздел 2) с указанием этапов (семестров) их освоения.

Результаты обучения вносят свой вклад в формирование различных компетенций, предусмотренных образовательной программой. В свою очередь, компетенции на разных уровнях категорий «знать», «уметь», «владеть» формируются модулями (разделами) дисциплины, а также различными дисциплинами образовательной программы.

2. ОПИСАНИЕ ПОКАЗАТЕЛЕЙ И КРИТЕРИЕВ ОЦЕНИВАНИЯ КОМПЕТЕНЦИЙ НА РАЗЛИЧНЫХ ЭТАПАХ ИХ ФОРМИРОВАНИЯ, ОПИСАНИЕ ШКАЛ ОЦЕНИВАНИЯ

Фонд оценочных средств (ФОС) предусматривает:

- анализ комплекса **показателей** - дескрипторов освоения компетенций в виде результатов обучения, которые студент может продемонстрировать (см. табл.1). Для контроля достижения каждого из них должны быть предусмотрены оценочные средства в виде вопросов, заданий и т.д.;
- обозначение **критериев** - правил принятия решения по оценке достигнутых результатов обучения и сформированности компетенций.

В качестве шкалы оценивания принимается 100-бальная система с выделением с градацией оценок в соответствии с Положением о текущем контроле успеваемости и промежуточной аттестации студентов МГТУ им. Н.Э. Баумана:

Рейтинг	**Оценка на экзамене**
85 – 100	отлично
71 – 84	хорошо
60 – 70	удовлетворительно
0 – 59	неудовлетворительно

Показатели достижения планируемых результатов обучения и критерии их оценивания на разных уровнях формирования компетенций приведены в таблице 1.

.

Таблица 1. Показатели достижения индикаторов компетенции

1	2	3	4
Компетенция: код по ФГОС 3++, формулировка	**Индикаторы**	**Этап**	**Наименование оценочного средства**
УК – 6 (09.04.01) Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки	ЗНАТЬ - методики самооценки, самоконтроля и саморазвития, в том числе с использованием подходов здоровьесбережения УМЕТЬ - решать задачи собственного личностного и профессионального развития, определять и реализовывать приоритеты совершенствования собственной деятельности	1	Рубежные контроли Защита лабораторных работ Домашние задания Экзамен
ПК – 4 Способен руководить проектами по созданию, поддержке и использованию системы искусственного интеллекта на основе нейросетевых моделей и методов	**ПК-4.1. Руководит работами по оценке и выбору моделей искусственных нейронных сетей и инструментальных средств для решения поставленной задачи** ЗНАТЬ - возможности современных инструментальных средств и систем программирования для решения задач машинного обучения.	1	Рубежные контроли Защита лабораторных работ Домашние задания Экзамен

Критерии оценки результатов обучения при различных видах контрольных мероприятий приведены в таблицах:

Критерии оценивания результатов рубежного контроля (РК):	
Критерии начисления баллов	**Оценка**
Ответы содержательные, уверенные и четкие; показано свободное владение материалом различной степени сложности	*От_85_ до_100_% от максимального количества баллов за РК*
Твердо усвоен основной материал; ответы удовлетворяют требованиям, установленным для максимальной оценки, но при этом допускаются негрубые ошибки; делаются несущественные пропуски при изложении фактического материала.	*От_71_ до_84_ % от максимального количества баллов за РК*
Обучаемый знает и понимает основной материал программы, основные темы, но в усвоении материала имеются пробелы; излагает его упрощенно, с ошибками и затруднениями; изложение теоретического материала приводится с ошибками.	*От_60_ до_70_ % от максимального количества баллов за РК*
Отказ от ответа или отсутствие минимальных знаний по дисциплине; присутствуют грубые ошибки в ответе.	*От_0_ до_59_ % от максимального количества баллов за РК*

Критерии оценивания домашнего задания (реферата, доклада):

От 9 до 10 баллов: глубокое и хорошо аргументированное обоснование темы; четкая формулировка и понимание изучаемой проблемы; широкое и правильное использование относящейся к теме литературы и примененных аналитических методов; содержание исследования и ход защиты указывают на наличие навыков работы студента в данной области; оформление работы хорошее с наличием расширенной библиографии; защита реферата (выступление с докладом) показала высокий уровень профессиональной подготовленности студента;

От 7 до 8 баллов: аргументированное обоснование темы; четкая формулировка и понимание изучаемой проблемы; использование ограниченного, но достаточного для проведения исследования количества источников; работа основана на среднем по глубине анализе изучаемой проблемы и при этом сделано незначительное число обобщений; содержание исследования и ход защиты (выступление с докладом) указывают на наличие практических навыков работы студента в данной области; реферат (доклад) хорошо оформлен с наличием необходимой библиографии; ход защиты реферата (выступления с докладом) показал достаточную научную и профессиональную подготовку студента;

6 баллов: достаточное обоснование выбранной темы, но отсутствует глубокое понимание рассматриваемой проблемы; в библиографии преобладают ссылки на стандартные литературные источники; труды, необходимые для всестороннего изучения проблемы, использованы в ограниченном объеме; заметна нехватка компетентности студента в данной области знаний; оформление реферата (доклада) содержит небрежности; защита реферата (выступление с докладом) показала удовлетворительную профессиональную подготовку студента;

От 0 до 5 баллов: тема реферата (доклада) представлена в общем виде; ограниченное число использованных литературных источников; шаблонное изложение материала; суждения по исследуемой проблеме не всегда компетентны; неточности и неверные выводы по рассматриваемой литературе; оформление реферата (доклада) с элементами заметных отступлений от общих требований; во время защиты (выступления с докладом) студентом проявлена ограниченная профессиональная эрудиция.

Критерии оценивания лабораторных работ в модуле (ЛР):

Степень выполнения лабораторных работ оценивается по степени выполнения каждой лабораторной работы в модуле и суммированием набранных процентов.

Степень выполнения каждой ЛР оценивается по шкале:

Критерии начисления баллов	**Оценка**
за правильно работающую программу и успешно защищенную; полный отчёт по ЛР (в соответствии с методическими указаниями)	*От 85 до 100 % , делённое на количество ЛР в модуле*
за работающую программу при наличии мелких некорректностей и самостоятельного их исправления в процессе защиты; полный отчёт по ЛР (в соответствии с методическими указаниями)	*От 71 до 84 % , делённое на количество ЛР в модуле*
за работающую программу при наличии мелких некорректностей и самостоятельного их исправления в процессе защиты; отчёт по ЛР оформлен с недочетами.	*От 60 до 70 % , делённое на количество ЛР в модуле*
за неработающую программу при наличии некорректностей и их неисправлении по подсказке преподавателя в процессе защиты; студент представил недостаточно полный (в соответствии с методическими указаниями) отчёт.	*От_0_ до_59_ %, делённое на количество ЛР в модуле*
Набранные проценты за все ЛР суммируются и пересчитываются в баллы. Итоговая оценка за все ЛР в модуле 1, в модуле 2:	
От 85 до 100 %	От 13 до 15 баллов
От 71 до 84 %	От 11 до 12 баллов
От 60 до 70 %	От 9 до 10 баллов
От 0 до 59 %	От 0 до 8 баллов

Критерии оценивания на экзамене:

От 25 до 30 баллов и/или «отлично» в: студент глубоко и полно владеет содержанием учебного материала и понятийным аппаратом; умеет связывать теорию с практикой, иллюстрировать примерами, фактами, данными научных исследований; осуществляет межпредметные связи, предложения, выводы; логично, четко и ясно излагает ответы на поставленные вопросы; умеет обосновывать свои суждения и профессионально-личностную позицию по излагаемому вопросу; ответ носит самостоятельный характер.

От 21 до 24 баллов и/или «хорошо»: ответ студента соответствует указанным выше критериям, но в содержании имеют место отдельные неточности (несущественные ошибки) при изложении теоретического и практического материала; ответ отличается меньшей обстоятельностью, глубиной, обоснованностью и полнотой; однако допущенные ошибки исправляются самим студентом после дополнительных вопросов экзаменатора.

От 18 до 20 баллов и/или «удовлетворительно»: студент обнаруживает знание и понимание основных положений учебного материала, но излагает его неполно, непоследовательно, допускает неточности и существенные ошибки в определении понятий, формулировке положений; при аргументации ответа студент не опирается на основные положения исследовательских документов; не применяет теоретические знания для объяснения эмпирических фактов и явлений, не обосновывает свои суждения; имеет место нарушение логики изложения; в целом ответ отличается низким уровнем самостоятельности, не содержит собственной профессионально-личностной позиции.

От 0 до 17 баллов и/или «неудовлетворительно»: студент имеет разрозненные, бессистемные знания; не умеет выделять главное и второстепенное; в ответе допускаются

ошибки в определении понятий, формулировке теоретических положений, искажающие их смысл; студент не ориентируется в нормативно-концептуальных, программно-методических, исследовательских материалах, беспорядочно и неуверенно излагает материал; не умеет соединять теоретические положения с педагогической практикой; не умеет применять знания для объяснения эмпирических фактов, не устанавливает межпредметные связи.

Использование показателей и критериев оценивания компетенций на различных этапах их формирования совместно со шкалой балльно-рейтинговой системы позволяет формировать результаты обучения по модулям.

Оценка результатов обучения

Неделя	**Номер и название модуля**	**Формы контроля**	**Баллы (мин/ макс)**
1 семестр			
7	1. Базовые понятия систем обработки больших данных	Рубежный контроль №1	6 / 10
		Защита лабораторных работ	9 / 15
		Домашнее задание №1	6 / 10
		ИТОГО	**21 / 35**
11	2. NoSQL- и SQL-системы баз данных, использующие шардирование	Рубежный контроль №2	6 / 10
		Защита лабораторных работ	9 / 15
		Домашнее задание №2	6 / 10
		ИТОГО	**21 / 35**
	3. Экзамен	-	18 / 30
		ИТОГО за семестр	**60 / 100**

3. ТИПОВЫЕ КОНТРОЛЬНЫЕ ЗАДАНИЯ ИЛИ ИНЫЕ МАТЕРИАЛЫ, НЕОБХОДИМЫЕ ДЛЯ ОЦЕНКИ ЗНАНИЙ, УМЕНИЙ, НАВЫКОВ И (ИЛИ) ОПЫТА ДЕЯТЕЛЬНОСТИ, ХАРАКТЕРИЗУЮЩИХ ЭТАПЫ ФОРМИРОВАНИЯ КОМПЕТЕНЦИЙ В ПРОЦЕССЕ ОСВОЕНИЯ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

ФОС по дисциплине содержит следующие оценочные средства, позволяющие оценить знания, умения и уровень приобретенных компетенций при текущем и промежуточном контроле, разбитые по модулям дисциплины:

- перечень типовых вопросов для защиты лабораторных работ;
- перечни типовых тем для домашних заданий (реферат, доклад);
- перечень вопросов для рубежных контролей;
- примеры вариантов рубежных контролей;
- перечень экзаменационных вопросов;
- макет билета к экзамену.

Средства для оценки различных уровней формирования компетенций по категориям знать, уметь, владеть обеспечивают реализацию следующих основных принципов контроля, таких, как объективность и независимость, практико-ориентированность, междисциплинарность.

С учетом этого, контрольные вопросы, задания, задачи, входящие а ФОС, для различных категорий и уровней освоения компетенций имеют вид:

Уровень ЗНАТЬ

Дескрипторы	Пример задания из оценочного средства
методики самооценки, самоконтроля и саморазвития, в том числе с использованием подходов здоровьесбережения	1. Почему на практике чаще используются коммерческие дистрибутивы Hadoop при наличии Open Source дистрибутива? 2. В каких случаях компании из не IT-бизнеса (клиенту) целесообразно использовать OpenStack в качестве компонента IT-инфраструктуры? 3. Какую роль играют краулеры в работе Search Engines? Могут ли Search Engines работать без краулеров? 4. Какие технические решения и подходы должны применять разработчики традиционных систем управления базами данных, чтобы противостоять давлению инструментов и технологий обработки больших данных, постепенно расширяя границы применения своих СУБД за счёт увеличения предельных объёмов обрабатываемых данных?
возможности современных инструментальных средств и систем программирования для решения задач машинного обучения	

Уровень УМЕТЬ

Дескрипторы	Пример задания из оценочного средства
решать задачи собственного личностного и профессионального развития, определять и реализовывать приоритеты совершенствования собственной деятельности	1. Как осуществляется настройка производительности Hadoop? 2. Как можно и можно ли использовать Oozie, не применяя MapReduce? 3. Предложите MapReduce схему получения списка 100 человек, получивших в этом году самый большой доход, по хранимому в HDFS файлу налоговых деклараций всех

Дескрипторы	Пример задания из оценочного средства
решать задачи собственного личностного и профессионального развития, определять и реализовывать приоритеты совершенствования собственной деятельности	жителей страны. Ключом в файле является номер ИНН. Что дополнительно необходимо, чтобы получить список 100 самых богатых человек. Предложите соответствующее решение средствами MapReduce / YARN. 4. Сопоставьте возможности YARN и Mesos, укажите их сильные и слабые стороны. 5. Сопоставьте Hadoop 2.0 и его окружение со Spark и его окружением. Когда выгодно использовать один стек технологий, а когда другой? 6. Сопоставьте Storm и Spark Streaming. Приведите их сильные и слабые стороны.

4. МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ, ОПРЕДЕЛЯЮЩИЕ ПРОЦЕДУРЫ ОЦЕНИВАНИЯ ЗНАНИЙ, УМЕНИЙ, НАВЫКОВ И (ИЛИ) ОПЫТА ДЕЯТЕЛЬНОСТИ, ХАРАКТЕРИЗУЮЩИХ ЭТАПЫ ФОРМИРОВАНИЯ КОМПЕТЕНЦИЙ.

4.1. Примеры (макеты) методических материалов, определяющих процедуры оценивания знаний, умений, навыков и (или) опыта деятельности.

Наименование оценочного средства	Краткая характеристика оценочного средства	Представление оценочного средства в фонде
Рубежный контроль	Средство проверки освоения уровней «знать», «уметь» компетенций ФГОС 3++	Перечень вопросов для рубежных контролей; примеры вариантов рубежных контролей.
Защита лабораторных работ	Средство проверки освоения уровней «знать», «уметь» компетенций ФГОС 3++	Перечень типовых вопросов для защиты лабораторных работ.
Домашнее задание (реферат, доклад, презентация)	Средство проверки освоения уровней «знать» компетенций ФГОС 3++	Перечни типовых тем для домашних заданий (реферат, доклад)
Экзамен	Средство проверки освоения уровней «знать» компетенций ФГОС 3++	Перечень вопросов к экзамену; макет экзаменационного билета.

Перечень вопросов для рубежных контролей
Кафедра «Прикладной информатики и математики»
по дисциплине «Технологии обработки больших данных»

Рубежный контроль №1

1. Какие разделы появятся в технических дисциплинах в недалёком будущем под влиянием развития технологий обработки больших данных? (не практические задачи, для решения которых будут использоваться эти технологии, и не новые приложения, а именно разделы технических дисциплин).
2. Какие разделы появятся в экономических дисциплинах в недалёком будущем под влиянием развития технологий обработки больших данных? (не практические задачи, для решения которых будут использоваться эти технологии, и не новые приложения, а именно разделы экономических дисциплин).
3. Какие разделы появятся в естественно-научных дисциплинах в недалёком будущем под влиянием развития технологий обработки больших данных? (не практические задачи, для решения которых будут использоваться эти технологии, и не новые приложения, а именно разделы естественно-научных дисциплин).
4. Какие технические решения и подходы должны применять разработчики традиционных систем управления базами данных, чтобы противостоять давлению инструментов и технологий обработки больших данных, постепенно расширяя границы применения своих СУБД за счёт увеличения предельных объёмов обрабатываемых данных?

5. Какие схемы обработки данных можно реализовать в YARN, которые нельзя реализовать в Hadoop 1.0? Приведите примеры таких схем. Опишите механизмы YARN, делающие возможным их реализацию.
6. Как связаны MapReduce и HDFS?
7. Предложите MapReduce схему получения списка 100 человек, получивших в этом году самый большой доход, по хранимому в HDFS файлу налоговых деклараций всех жителей страны. Ключом в файле является номер ИНН. Что дополнительно необходимо, чтобы получить список 100 самых богатых человек. Предложите соответствующее решение средствами MapReduce / YARN.
8. Предложите схему обработки результатов голосования в Госдуму в сентябре этого года с помощью MapReduce / YARN. Голосование будет проводиться по смешанной системе, т.е. одновременно и по партийным спискам, и по одномандатным округам. Каждый избиратель будет голосовать по двум бюллетеням. По партийным спискам будет избираться ровно половина численного состава Госдумы – 225 человек, и столько же – по одномандатным округам. Партийный список допускается до распределения депутатских мандатов в том случае, если за него проголосовало более 5 % от числа всех избирателей, принявших участие в голосовании. По одномандатному избирательному округу победителем признается кандидат, получивший относительное большинство голосов избирателей округа. Если в одномандатном округе проголосует менее 25% от общего числа избирателей округа, выборы признаются недействительными и назначается повторное голосование. В обоих наборах данных, полученных в результате голосования, в качестве ключа будет использоваться серия и номер паспорта избирателя.
9. В каких случаях компании из не IT-бизнеса (клиенту) целесообразно использовать OpenStack в качестве компонента IT-инфраструктуры?
10. В каких случаях компании из IT-бизнеса целесообразно использовать OpenStack для решения собственных задач?
11. Какие проекты (подсистемы) OpenStack вероятнее всего потребуют доработки при внесении изменений в проект (подсистему) Sahara т.к. они теснее зависимы?
12. Возможны ли 3 уровня размещения данных в системе хранения, реализующей Hierarchical Storage Management (HSM)? Если да, то приведите пример. А 4? А больше?
13. Исходя из каких критериев компания должна выбрать используемую распределённую файловую систему?
14. В чём заключается принципиальное отличие файловых систем для HPC-кластеров (Lustre и др.) от файловых систем для кластеров Big Data?
15. В каких случаях компании из не IT-бизнеса (клиенту) целесообразно использовать одновременно несколько разных распределённых файловых систем?
16. Как повлияет распространение использования NVM на распределённые файловые системы? А как на локальные?
17. Какая из рассматривавшихся в курсе распределённых файловых систем была разработана в рамках работы над диссертацией? Какие положения были вынесены на защиту в этой диссертации?
18. Приведите пример, когда может быть целесообразно в настоящее время использовать MPI для организации распределённой обработки данных?
19. Сравните технологи MPI и Hadoop. В чём преимущества и недостатки каждой из них по отношению друг к другу?
20. Почему на практике чаще используются коммерческие дистрибутивы Hadoop при наличии Open Source дистрибутива?
21. Как обеспечивается функциональное соответствие между большим числом коммерческих дистрибутивов Hadoop и Open Source дистрибутивом по мере появления новых версий?
22. Как осуществляется настройка производительности Hadoop?
23. Как используется BigTop в процессе обновления Hadoop?

24. Приведите несколько способов реструктуризации большого набора данных в HDFS (удаление или перестановка порядка полей, изменение форматов полей). В чём состоят преимущества и недостатки каждого из приведенных способов?
25. Каким основным требованиям должны отвечать программы и алгоритмы архивации, используемые при обработке больших данных?
26. Возможны ли в классическом Hadoop следующие схемы обработки данных MapMapRedce, MapReduceReduce и более сложные комбинации в рамках одного Job?
27. Какие ещё существуют менеджеры ресурсов кластера помимо YARN, решающие примерно те же задачи? В чём их отличия от JARN?
28. Какие фреймворки реализованы на базе YARN кроме MapReduce?
29. Pig и Oozie используются клиентами для автоматизации обработки данных с помощью Hadoop. В каких случаях целесообразно использовать Pig, а в каких Oozie? Когда желательно применять оба продукта совместно?
30. Как можно и можно ли использовать Oozie, не применяя MapReduce?
31. Какие категории пользователей имеются у Tez? Как они его используют?
32. Какие из целей, ставившихся при создании Spark, были достигнуты полностью, а какие лишь частично?
33. В каких случаях наиболее важным преимуществом Spark над Hadoop 2.X является скорость выполнения заданий, а в каких – общее время подготовки и выполнения заданий?
34. Сопоставьте возможности YARN и Mesos, укажите их сильные и слабые стороны.
35. Какие проблемы сдерживают распространение решений Big Data, независимо от развития технологических возможностей и перехода с Hadoop на Spark и последующие поколения технологий?
36. Сопоставьте Hadoop 2.0 и его окружение со Spark и его окружением. Когда выгодно использовать один стек технологий, а когда другой?
37. Сопоставьте Spark и его окружение с SAP HANA и её окружением. Когда выгодно использовать один стек технологий, а когда другой?
38. Сопоставьте Flink и его окружение с SAP HANA и её окружением. Когда выгодно использовать один стек технологий, а когда другой?
39. Назначение и функции Modeler’a в SAP HANA. Можно ли решать какие-нибудь задачи в SAP HAMA без использования Modeler’a?
40. Какие функциональные возможности Modeler’a в SAP HANA лучше всего подходят для создания экспертных систем на базе платформы SAP HANA? Почему?
41. Сильные и слабые стороны SAP LT. Типичные варианты его применения.
42. Сильные и слабые стороны SAP Data Services (DS). Типичные варианты его применения.
43. Сильные и слабые стороны Smart Data Access. Типичные варианты его применения.
44. Сильные и слабые стороны SAP HANA Direct Extractor Connection (DXC). Типичные варианты его применения.
45. Предложите варианты загрузки данных в SAP HANA, не описанные в курсе HA350.
46. Какой проект пытается преодолеть недостатки Spark и Flink и совместить их сильные стороны. В чём заключается суть этого проекта?
47. Чем Kafka отличается от традиционных систем организации и ведения очередей?
48. Как строится эшелонированная система агентов Flume? Зачем? Какие задачи она решает? Как она отображается на структуре серверов?
49. Сопоставьте Storm и Spark Streaming. Приведите их сильные и слабые стороны.
50. Сопоставьте Storm и Samza. Приведите их сильные и слабые стороны.
51. Сопоставьте Storm и Flink. Приведите их сильные и слабые стороны.
52. Сопоставьте Spark Streaming и Samza. Приведите их сильные и слабые стороны.
53. Сопоставьте Spark и Flink. Приведите их сильные и слабые стороны.
54. Сопоставьте Samza и Flink. Приведите их сильные и слабые стороны.
55. Сопоставьте Spark Streaming и Trident. Приведите их сильные и слабые стороны.

56. Какие существуют возможности для формирования декларативных запросов, определяющих обобщённые данные, которые необходимо получить в ходе обработки потока данных? Где и каким образом это можно реализовать?

Рубежный контроль №2

1. Основные модели данных NoSQL СУБД. Применение каждой из моделей.
2. Типы NoSQL СУБД. Применение каждого из типов.
3. Чем обусловлена востребованность NoSQL СУБД?
4. Как блокируются ресурсы средствами Zookeeper? Как можно сделать этими средствами барьер?
5. Чем отличается модель данных Zookeeper от большинства других СУБД класса Key-Value?
6. Чем отличается организация работы кластера Zookeeper от большинства других СУБД класса Key-Value?
7. Сопоставьте возможности нескольких суперстолбцовых NoSQL СУБД.
8. Как организовано взаимодействие HBase и Zookeeper?
9. Сопоставьте возможности суперстолбцовых NoSQL СУБД, традиционных SQL СУБД и Big Data SQL СУБД.
10. Особенности архитектуры и организации данных СУБД Accumulo
11. Особенности архитектуры и организации данных СУБД Cassandra.
12. Основные отличия между CQL и SQL.
13. Сопоставьте возможности нескольких документарных NoSQL СУБД.
14. Возможности поиска и манипулирования данными MongoDB.
15. Основные отличия между N1QL и SQL.
16. Сопоставьте возможности нескольких NoSQL СУБД класса Key-Value.
17. Особенности архитектуры и модели данных NoSQL СУБД Redis.
18. Возможности поиска и манипулирования данными СУБД Redis.
19. Особенности архитектуры и организации данных СУБД Kudu. Чем обусловлена более высокая скорость обработки данных в СУБД Kudu?
20. Каковы внутренние отличия SQL и NoSQL СУБД? Какие проблемы не могут быть решены средствами NoSQL СУБД и заставляют использовать SQL СУБД?
21. Какие элементы HiveQL обеспечивают более высокую скорость работы с большими данными по сравнению со стандартным SQL?
22. Почему Impala долгое время опережала по скорости работы Hive? Каким образом команда Hive пытается догнать Impala?
23. Сопоставьте возможности Phoenix и Drill. Приведите их сильные и слабые стороны.
24. В чём состоят преимущества и недостатки подхода, выбранного командами Phoenix, Drill и других аналогов к созданию SQL СУБД?
25. Сопоставьте возможности Impala и HAWQ. Приведите их сильные и слабые стороны. Когда целесообразнее использовать одну, а когда другую?
26. Чем отличаются подходы в создании и развитии SQL СУБД Impala и Vertica? Как СУБД Impala будет развиваться, чтобы достигнуть такого же состояния, что и Vertica?
27. В чём состоят преимущества и недостатки Spark SQL?
28. Чем отличаются подходы в создании и развитии SQL СУБД HP Vertica и IBM Big SQL?
29. Сопоставьте возможности SQL СУБД в составе платформ SAP HANA и Spark. Приведите их сильные и слабые стороны.
30. В чём состоят преимущества и недостатки SAP HANA Vora?
31. Почему SAP потребовалось вывести на рынок две SQL СУБД для работы с большими данными – в составе платформы SAP HANA и SAP HANA Vora? Как они могут использоваться вместе и по отдельности?
32. Сопоставьте возможности SAP HANA Vora и IBM Big SQL. Приведите их сильные и слабые стороны.

33. Сопоставьте возможности SQL СУБД, присутствующих в составе платформ Big Data (Spark, SAP HANA, Flink и др.), и возможности независимых Big Data SQL СУБД (HP Vertica, IBM Big SQL и др.). Когда целесообразнее использовать первые, а когда вторые?
34. В чём состоят три революции, имевшие место в области СУБД?
35. В чём заключаются сходства и различия процессов развития СУБД перед первой и перед второй революцией в этой области, а также в ходе каждой из них?
36. В чём заключаются сходства и различия процессов развития СУБД перед первой и перед третьей революцией в этой области, а также в ходе каждой из них?
37. В чём состоят проблемы, нерешённые в ходе третьей революции в области СУБД?
38. Чем обусловлен взрывной рост интереса в последние 2-3 года к графовым СУБД?
39. Сопоставьте возможности СУБД Neo4j, OrientDB и Titan. Приведите их сильные и слабые стороны.
40. В чём состоят различия между графовыми СУБД и Graph Engines?
41. Как соотносятся между собой между собой Giraph и Pregel?
42. В чём заключается новизна решений Surfer, GBASE, Signal/Collect и Mizan?
43. В чём состоят отличия PEGASUS от Pregel, её клонов и потомков?
44. Как проходила эволюция продуктов GraphLab? В чём новизна решений, реализуемых на каждом этапе?
45. Преимущества и недостатки Apache Spark GraphX по сравнению с независимо развиваемыми Graph Engines?
46. Каким образом Spark GraphX может использоваться для графов, хранимых в графовых СУБД?
47. Что необходимо добавить в графовую СУБД, чтобы она начала выполнять функции Graph Engine? Что необходимо добавить в Graph Engine, чтобы он начал выполнять функции графовой СУБД?
48. Сопоставьте современное состояние продуктов, сочетающих в себе возможности графовых СУБД и Graph Engines, с отдельно развиваемыми графовыми СУБД и отдельно развиваемыми Graph Engines.
49. Сопоставьте между собой SAP HANA Graph Engine и Microsoft Graph Engine, покажите их сильные и слабые стороны по отношению друг к другу.
50. Основные принципы работы и общие черты Search Engines.
51. Какую роль играют краулеры в работе Search Engines? Могут ли Search Engines работать без краулеров?
52. Общий алгоритм работы Nutch.
53. Сопоставьте возможности Solr и Elasticsearch. Приведите их сильные и слабые стороны.
54. Роль Search Engines в корпоративных системах масштаба предприятия.

Примеры вариантов рубежных контролей

Кафедра «Прикладной информатики и математики
по дисциплине «Технологии обработки больших данных»

Пример варианта рубежного контроля №1

1. В чём заключается принципиальное отличие файловых систем для HPC-кластеров (Lustre и др.) от файловых систем для кластеров Big Data?
2. Сильные и слабые стороны Smart Data Access. Типичные варианты его применения.

Пример варианта рубежного контроля №2

1. Как блокируются ресурсы средствами Zookeeper? Как можно сделать этими средствами барьер?
2. Почему Impala долгое время опережала по скорости работы Hive? Каким образом команда Hive пытается догнать Impala?

Составитель _____ Абдеева Н..А.
(подпись)

Перечень типовых вопросов для защиты лабораторных работ
Кафедра «Прикладной информатики и математики»
по дисциплине «Технологии обработки больших данных»

Лабораторная работа №1. Знакомство с Hadoop.
Краткое описание: Запуск виртуальной машины с гостевой ОС Ubuntu. Использование этой машины для развёртывания и настройки Apache Hadoop в SingleNode режиме. Проверка работоспособности Hadoop с помощью тестовых программ. Далее студенты объединяются в группы по 3-4 человека и выполняют настройку Multinode кластера Hadoop, а затем проверяют работоспособность настроенного кластера, выполняя тестовые программы.
Типовые вопросы для защиты лабораторной работы №1:

1. Какие схемы обработки данных можно реализовать в YARN, которые нельзя реализовать в Hadoop 1.0? Приведите примеры таких схем. Опишите механизмы YARN, делающие возможным их реализацию.
2. Как связаны MapReduce и HDFS?
3. Предложите MapReduce схему получения списка 100 человек, получивших в этом году самый большой доход, по хранимому в HDFS файлу налоговых деклараций всех жителей страны. Ключом в файле является номер ИНН. Что дополнительно необходимо, чтобы получить список 100 самых богатых человек. Предложите соответствующее решение средствами MapReduce / YARN.
4. Предложите схему обработки результатов голосования в Госдуму в сентябре этого года с помощью MapReduce / YARN. Голосование будет проводиться по смешанной системе, т.е. одновременно и по партийным спискам, и по одномандатным округам. Каждый избиратель будет голосовать по двум бюллетеням. По партийным спискам будет избираться ровно половина численного состава Госдумы – 225 человек, и столько же – по одномандатным округам. Партийный список допускается до распределения депутатских мандатов в том случае, если за него проголосовало более 5 % от числа всех избирателей, принявших участие в голосовании. По одномандатному избирательному округу победителем признается кандидат, получивший относительное большинство голосов избирателей округа. Если в одномандатном округе проголосует менее 25% от общего числа избирателей округа, выборы признаются недействительными и назначается повторное голосование. В обоих наборах данных, полученных в результате голосования, в качестве ключа будет использоваться серия и номер паспорта избирателя.
5. Почему на практике чаще используются коммерческие дистрибутивы Hadoop при наличии Open Source дистрибутива?
6. Как обеспечивается функциональное соответствие между большим числом коммерческих дистрибутивов Hadoop и Open Source дистрибутивом по мере появления новых версий?
7. Как осуществляется настройка производительности Hadoop?

8. Приведите несколько способов реструктуризации большого набора данных в HDFS (удаление или перестановка порядка полей, изменение форматов полей). В чём состоят преимущества и недостатки каждого из приведенных способов?
9. Возможны ли в классическом Hadoop следующие схемы обработки данных MapMapRedce, MapReduceReduce и более сложные комбинации в рамках одного Job?
10. Какие ещё существуют менеджеры ресурсов кластера помимо YARN, решающие примерно те же задачи? В чём их отличия от JARN?

Лабораторная работа №2. Знакомство со Spark.
Краткое описание: Разворачивание Spark с помощью менеджера кластера. Загрузка в Spark набора данных. Выполнить 2 простых запроса по обработке загруженного набора данных. Разработать сервис рекомендаций на основе набора данных.
Типовые вопросы для защиты лабораторной работы №2:
1. Основные компоненты Spark?
2. В каких случаях наиболее важным преимуществом Spark над Hadoop 2.X является скорость выполнения заданий, а в каких – общее время подготовки и выполнения заданий?
3. Сопоставьте возможности Spark поверх YARN и поверх Mesos, укажите преимущества и слабые стороны.
4. Какие проблемы сдерживают распространение решений Big Data, независимо от развития технологических возможностей и перехода с Hadoop на Spark и последующие поколения технологий?
5. Сопоставьте Hadoop 2.0 и его окружение со Spark и его окружением. Когда выгодно использовать один стек технологий, а когда другой?
6. Сопоставьте Spark и его окружение с SAP HANA и её окружением. Когда выгодно использовать один стек технологий, а когда другой?
7. Представление данных в Spark – RDD, Dataset и Dataframe, в чём состоят основные отличия?
8. Эволюция функциональности Spark в последних версиях?
9. Экосистема вокруг Spark?
10. Как Spark работает на кластере под управлением Kubernetes?

Лабораторная работа №3. Знакомство с Kubernetes и контейнеризацией вычислений.
Краткое описание: Регистрация триального экаунта в Google Cloud, получение на счёт денежных средств для работы с Google Cloud. Резервирование кластера и передача его под управление Kubernetes и Docker. Развёртывание на этом кластере Sparc под Kubernetes и Docker. Выполнение Spark Job. Перенос одной из нод Spark контейнером на другую ВМ. Повторное выполнение Spark Job.
Типовые вопросы для защиты лабораторной работы №3:
1. Сопоставьте контейнеры и виртуальные машины. В чём состоят преимущества и в чём недостатки каждого подхода к организации вычислений?
2. Что представляет собой Docker?
3. Как работает система оркестрации контейнеров?
4. В чём состоит отличие оркестрации контейнеров от управления контейнерами?
5. Что значит контейнеризованное приложение?
6. Как связана контейнеризация с микросервисной архитектурой?
7. Как работает приложение, имеющее Serverless архитектуру?
8. Основные возможности Kubernetes?
9. Альтернативы Docker?
10. Альтернативы Kubernetes?
11. Экосистема Kubernetes?

Лабораторная работа №4. Сложная обработка больших данных в экосистеме Hadoop.
Краткое описание: Разворачивание Hadoop, Pig и Hive с помощью менеджера кластера. Загрузка в Hadoop набора данных. Разработка и выполнение трёх разных запросов в Pig и таких же по смыслу запросов в Hive. Сопоставление времени разработки и времени выполнения в обоих инструментах.
Типовые вопросы для защиты лабораторной работы №4:

1. Pig и Oozie используются клиентами для автоматизации обработки данных с помощью Hadoop. В каких случаях целесообразно использовать Pig, а в каких Oozie? Когда желательно применять оба продукта совместно?
2. Как можно и можно ли использовать Oozie, не применяя MapReduce?
3. Какие категории пользователей имеются у Tez? Как они его используют?
4. Чем Kafka отличается от традиционных систем организации и ведения очередей?
5. Как строится эшелонированная система агентов Flume? Зачем? Какие задачи она решает? Как она отображается на структуре серверов?
6. Сопоставьте Storm и Spark Streaming. Приведите их сильные и слабые стороны.
7. Сопоставьте Storm и Samza. Приведите их сильные и слабые стороны.
8. Сопоставьте Storm и Flink. Приведите их сильные и слабые стороны.
9. Сопоставьте Spark Streaming и Samza. Приведите их сильные и слабые стороны.
10. Сопоставьте Spark и Flink. Приведите их сильные и слабые стороны.
11. Сопоставьте Samza и Flink. Приведите их сильные и слабые стороны.
12. Сопоставьте Spark Streaming и Trident. Приведите их сильные и слабые стороны.
13. Какие существуют возможности для формирования декларативных запросов, определяющих обобщённые данные, которые необходимо получить в ходе обработки потока данных? Где и каким образом это можно реализовать?

Лабораторная работа №5. Знакомство с SAP HANA.
Краткое описание: Установка SAP HANA Studio. Подключение к облачной SAP HANA. Установка полномочий пользователя. Выполнение примера по созданию двух таблиц и построению их объединения (Inner Join). Активация SAP HANA Graph. Формирование с его помощью двух простых графов в виде звезды м в виде кольца.
Типовые вопросы для защиты лабораторной работы №1:

1. Назначение и функции Modeler’a в SAP HANA. Можно ли решать какие-нибудь задачи в SAP HAMA без использования Modeler’a?
2. Какие функциональные возможности Modeler’a в SAP HANA лучше всего подходят для создания экспертных систем на базе платформы SAP HANA? Почему?
3. Сильные и слабые стороны SAP LT. Типичные варианты его применения.
4. Сильные и слабые стороны SAP Data Services (DS). Типичные варианты его применения.
5. Сильные и слабые стороны Smart Data Access. Типичные варианты его применения.
6. Сильные и слабые стороны SAP HANA Direct Extractor Connection (DXC). Типичные варианты его применения.
7. Основные модели данных NoSQL СУБД. Применение каждой из моделей.
8. Типы NoSQL СУБД. Применение каждого из типов.
9. Чем обусловлена востребованность NoSQL СУБД?

Лабораторная работа №6. Взаимодействие с SAP HANA через Web.
Краткое описание: **Часть 1.** Запуск XS-сервера в облачной SAP HANA. Разработка и отладка простой программы Hello Woirld. Разработка средствами XSJS OData-сервиса, обеспечивающего отображение таблицы из SAP HANA по запросам HTTP.Создание формы в HTML для ввода данных. Создание OData-сервиса для загрузки данных из формы HTML в таблицу SAP HANA.

Часть 2. Разработка CDS-сервиса для создания таблиц, View и схем в SAP HANA в процессе выполнения программ. Разработка сервиса для загрузки данных в SAP HANA из таблиц Excel и выгрузки данных в Excel из таблиц SAP HANA.

Типовые вопросы для защиты лабораторной работы №6:

1. Каковы внутренние отличия SQL и NoSQL СУБД? Какие проблемы не могут быть решены средствами NoSQL СУБД и заставляют использовать SQL СУБД?
2. Какие элементы HiveQL обеспечивают более высокую скорость работы с большими данными по сравнению со стандартным SQL?
3. Почему Impala долгое время опережала по скорости работы Hive? Каким образом команда Hive пытается догнать Impala?
4. Сопоставьте возможности Phoenix и Drill. Приведите их сильные и слабые стороны.
5. В чём состоят преимущества и недостатки подхода, выбранного командами Phoenix, Drill и других аналогов к созданию SQL СУБД?
6. Сопоставьте возможности Impala и HAWQ. Приведите их сильные и слабые стороны. Когда целесообразнее использовать одну, а когда другую?
7. Чем отличаются подходы в создании и развитии SQL СУБД Impala и Vertica? Как СУБД Impala будет развиваться, чтобы достигнуть такого же состояния, что и Vertica?
8. В чём состоят преимущества и недостатки Spark SQL?

Лабораторная работа №7. Обработка больших графов в Microsoft Graph Engine.

Краткое описание. Регистрация на странице Microsoft Graph Engine. Разработка с помощью Microsoft Graph Engine программы для обработки предоставляемого Microsoft графа ссылок, содержащихся в научных статьях. Программа должна рассчитывать индекс цитируемости каждой статьи.

Типовые вопросы для защиты лабораторной работы №7:

1. Чем обусловлен взрывной рост интереса в последние 2-3 года к графовым СУБД?
2. Сопоставьте возможности СУБД Neo4j, OrientDB и Titan. Приведите их сильные и слабые стороны.
3. В чём состоят различия между графовыми СУБД и Graph Engines?
4. Как соотносятся между собой между собой Giraph и Pregel?
5. В чём заключается новизна решений Surfer, GBASE, Signal/Collect и Mizan?
6. В чём состоят отличия PEGASUS от Pregel, её клонов и потомков?
7. Как проходила эволюция продуктов GraphLab? В чём новизна решений, реализуемых на каждом этапе?
8. Преимущества и недостатки Apache Spark GraphX по сравнению с независимо развиваемыми Graph Engines?
9. Каким образом Spark GraphX может использоваться для графов, хранимых в графовых СУБД?
10. Что необходимо добавить в графовую СУБД, чтобы она начала выполнять функции Graph Engine? Что необходимо добавить в Graph Engine, чтобы он начал выполнять функции графовой СУБД?
11. Сопоставьте современное состояние продуктов, сочетающих в себе возможности графовых СУБД и Graph Engines, с отдельно развиваемыми графовыми СУБД и отдельно развиваемыми Graph Engines.
12. Сопоставьте между собой SAP HANA Graph Engine и Microsoft Graph Engine, покажите их сильные и слабые стороны по отношению друг к другу.

Перечни типовых тем для домашних заданий (реферат, доклад)
Кафедра «Прикладной информатики и математики»
по дисциплине «Технологии обработки больших данных»

Домашнее задание №1

№	Тема
1	Понятие больших данных. Характеристики V. Задачи обработки больших данных в различных областях человеческой деятельности.
2	Источники и предпосылки возникновения технологий обработки больших данных. История возникновения, принципы работы и современное состояние Hadoop. Data Lake.
3	Иерархические системы хранения данных. NVM. Распределённые файловые системы. История возникновения распределённых файловых систем. GPFS, HDFS, Swift, Ceph, CFS. Обзор и перспективы развития распределённых файловых систем.
4	Проекты распределённых файловых систем экзобайтного масштаба. Перспективы развития распределённых файловых систем.
5	MPI.
6	Экосистема Hadoop. Дистрибутивы Hadoop. Интегрированные дистрибутивы Cloudera Enterprise, Hortonworks Data Platform, MapR Converged Data Platform.
7	Проприетарные и Open Source интегрированные платформы для обработки больших данных. IBM Bluemix, SAP HANA, Amazon Web Services, Microsoft Azure, Google Cloud Platform, Apache Spark, Apache Flink, Apache Beam.
8	Влияние технологий обработки больших данных на развитие дисциплин, связанных с моделированием. Численные методы, имитационное моделирование, теория игр, теория графов, исследование операций, оптимальное управление и др.
9	Цель использования и преимущества Лямбда-архитектуры.
10	Назначение, состав и функциональные возможности IBM Bluemix.
11	Экосистема и окружение SAP HANA. Шаблоны типовой архитектуры с использованием SAP HANA.
12	История развития, назначение и состав платформы OpenStack. Проект Sahara.
13	Основные элементы Hadoop. Режимы работы Hadoop. Обобщённая схема MapReduce.
14	Комбинирующие функции в MapReduce.
15	Архитектура HDFS. Преодоление сбоев и изоляция.
16	Интерфейс командной строки в Hadoop.
17	YARN.
18	MESOS.
19	Sqoop.
20	Ввод-вывод в Hadoop.
21	Apache Avro.
22	Администрирование Hadoop.

23	Сортировка и соединение наборов данных средствами MapReduce.
24	Parquet.
25	Pig и Pig Latin.
26	Oozie.
27	Hive.
28	Tez.
29	История развития и архитектура Spark. Программирование операций с RDD.
30	Работа в Spark с парами ключ-значение.
31	Загрузка и сохранение данных в Spark.
32	Spark Streaming.
33	Spark SQL.
34	Архитектура SAP HANA.
35	Графическое моделирование в SAP HANA.
36	SQL в SAP HANA.
37	Обработка текстов в SAP HANA.
38	Методы и инструментарий загрузки данных в SAP HANA.
39	Теоретические основы потоковой обработки данных.
40	SAP HANA Streaming.
41	Storm.
42	Kafka.
43	Flume.
44	Samza.
45	Flink.
46	Сопоставление систем потоковой обработки данных.
47	Равноправие поточной и пакетной обработки данных в Apache Beam.
48	Декларативные языки для обработки потоковых данных.
49	Trident поверх Storm.
50	SAP HANA CCL.

Домашнее задание №2

№	Тема
1	Причины появления систем NoSQL баз данных и их роль в развитии технологий обработки больших данных. 4 этапа развития СУБД и 3 революции, произошедшие в ходе этого развития.
2	Типы систем NoSQL баз данных и тенденции развития систем NoSQL баз данных.
3	Apache ZooKeeper.
4	HBase.

5	Hipertable.
6	Accumulo.
7	Cassandra.
8	MongoDB.
9	CouchDB как предшественник Couchbase.
10	RavenDB.
11	Redis.
12	Riak.
13	Kudu.
14	Задачи, которые решает Data Engineer? Роль технологий Big Data в комплексе его знаний и умений.
15	Архитектура корпоративных систем с использованием инструментов Big Data для интернет-компаний и для производственных компаний.
16	SQL vs NoSQL Database. История и тенденции развития SQL СУБД, способных обрабатывать большие данные.
17	Couchbase N1QL похожий на SQL.
18	HiveQL - почти как SQL.
19	Couchbase.
20	Apache Phoenix.
21	Apache Drill.
22	Cloudera Impala.
23	Apache HAWQ.
24	HP Vertica.
25	IBM BigSQL
26	SAP HANA Vora.
27	Интеграция традиционных SQL СУБД и средств обработки больших данных
28	Технология Polybase
29	Двухъярусные СУБД (NoSQL – SQL)
30	Мультимодельные СУБД
31	In-Memory СУБД
32	Транзакционные и аналитические СУБД с множественным шардированием
33	Общий обзор графовых СУБД.
34	Модели данных графовых СУБД
35	Neo4j.
36	OrientDB.
37	Titan.
38	Общий обзор Graph Engines.

39	Pregel.
40	Giraph.
41	Sufer, Signal/Collect и Mizan.
42	Семейство продуктов GraphLab.
43	PEGASUS.
44	Apache Spark GraphX.
45	Совмещение графовых СУБД и Graph Engine в одном продукте.
46	SAP HANA Graph.
47	Microsoft Graph Engine.
48	Обзор Search Engines.
49	Nutch и Solr как развитие Lucene.
50	ElasticSearch и Splunk.

Составитель _____ Абдеева Н. А.
(подпись)

Перечень вопросов к экзамену
Кафедра «Прикладной информатики и математики»
по дисциплине «Технологии обработки больших данных»

1. Какие разделы появятся в технических дисциплинах в недалёком будущем под влиянием развития технологий обработки больших данных? (не практические задачи, для решения которых будут использоваться эти технологии, и не новые приложения, а именно разделы технических дисциплин)
2. Какие разделы появятся в экономических дисциплинах в недалёком будущем под влиянием развития технологий обработки больших данных? (не практические задачи, для решения которых будут использоваться эти технологии, и не новые приложения, а именно разделы экономических дисциплин)
3. Какие разделы появятся в естественно-научных дисциплинах в недалёком будущем под влиянием развития технологий обработки больших данных? (не практические задачи, для решения которых будут использоваться эти технологии, и не новые приложения, а именно разделы естественно-научных дисциплин)
4. Какие технические решения и подходы должны применять разработчики традиционных систем управления базами данных, чтобы противостоять давлению инструментов и технологий обработки больших данных, постепенно расширяя границы применения своих СУБД за счёт увеличения предельных объёмов обрабатываемых данных?
5. Какие схемы обработки данных можно реализовать в YARN, которые нельзя реализовать в Hadoop 1.0? Приведите примеры таких схем. Опишите механизмы YARN, делающие возможным их реализацию.
6. Как связаны MapReduce и HDFS?
7. Предложите MapReduce схему получения списка 100 человек, получивших в этом году самый большой доход, по хранимому в HDFS файлу налоговых деклараций всех жителей страны. Ключом в файле является номер ИНН. Что дополнительно

необходимо, чтобы получить список 100 самых богатых человек. Предложите соответствующее решение средствами MapReduce / YARN.

8. Предложите схему обработки результатов голосования в Госдуму в сентябре этого года с помощью MapReduce / YARN. Голосование будет проводиться по смешанной системе, т.е. одновременно и по партийным спискам, и по одномандатным округам. Каждый избиратель будет голосовать по двум бюллетеням. По партийным спискам будет избираться ровно половина численного состава Госдумы – 225 человек, и столько же – по одномандатным округам. Партийный список допускается до распределения депутатских мандатов в том случае, если за него проголосовало более 5 % от числа всех избирателей, принявших участие в голосовании. По одномандатному избирательному округу победителем признается кандидат, получивший относительное большинство голосов избирателей округа. Если в одномандатном округе проголосует менее 25% от общего числа избирателей округа, выборы признаются недействительными и назначается повторное голосование. В обоих наборах данных, полученных в результате голосования, в качестве ключа будет использоваться серия и номер паспорта избирателя.
9. В каких случаях компании из не IT-бизнеса (клиенту) целесообразно использовать OpenStack в качестве компонента IT-инфраструктуры?
10. В каких случаях компании из IT-бизнеса целесообразно использовать OpenStack для решения собственных задач?
11. Какие проекты (подсистемы) OpenStack вероятнее всего потребуют доработки при внесении изменений в проект (подсистему) Sahara т.к. они теснее зависимы?
12. Возможны ли 3 уровня размещения данных в системе хранения, реализующей Hierarchical Storage Management (HSM)? Если да, то приведите пример. А 4? А больше?
13. Исходя из каких критериев компания должна выбрать используемую распределённую файловую систему?
14. В чём заключается принципиальное отличие файловых систем для HPC-кластеров (Lustre и др.) от файловых систем для кластеров Big Data?
15. В каких случаях компании из не IT-бизнеса (клиенту) целесообразно использовать одновременно несколько разных распределённых файловых систем?
16. Как повлияет распространение использования NVM на распределённые файловые системы? А как на локальные?
17. Какая из рассматривавшихся в курсе распределённых файловых систем была разработана в рамках работы над диссертацией? Какие положения были вынесены на защиту в этой диссертации?
18. Приведите пример, когда может быть целесообразно в настоящее время использовать MPI для организации распределённой обработки данных?
19. Сравните технологи MPI и Hadoop. В чём преимущества и недостатки каждой из них по отношению друг к другу?
20. Почему на практике чаще используются коммерческие дистрибутивы Hadoop при наличии Open Source дистрибутива?
21. Как обеспечивается функциональное соответствие между большим числом коммерческих дистрибутивов Hadoop и Open Source дистрибутивом по мере появления новых версий?
22. Как осуществляется настройка производительности Hadoop?
23. Как используется BigTop в процессе обновления Hadoop?
24. Приведите несколько способов реструктуризации большого набора данных в HDFS (удаление или перестановка порядка полей, изменение форматов полей). В чём состоят преимущества и недостатки каждого из приведенных способов?
25. Каким основным требованиям должны отвечать программы и алгоритмы архивации, используемые при обработке больших данных?

26. Возможны ли в классическом Hadoop следующие схемы обработки данных MapMapRedce, MapReduceReduce и более сложные комбинации в рамках одного Job?
27. Какие ещё существуют менеджеры ресурсов кластера помимо YARN, решающие примерно те же задачи? В чём их отличия от JARN?
28. Какие фреймворки реализованы на базе YARN кроме MapReduce?
29. Pig и Oozie используются клиентами для автоматизации обработки данных с помощью Hadoop. В каких случаях целесообразно использовать Pig, а в каких Oozie? Когда желательно применять оба продукта совместно?
30. Как можно и можно ли использовать Oozie, не применяя MapReduce?
31. Какие категории пользователей имеются у Tez? Как они его используют?
32. Какие из целей, ставившихся при создании Spark, были достигнуты полностью, а какие лишь частично?
33. В каких случаях наиболее важным преимуществом Spark над Hadoop 2.X является скорость выполнения заданий, а в каких – общее время подготовки и выполнения заданий?
34. Сопоставьте возможности YARN и Mesos, укажите их сильные и слабые стороны.
35. Какие проблемы сдерживают распространение решений Big Data, независимо от развития технологических возможностей и перехода с Hadoop на Spark и последующие поколения технологий?
36. Сопоставьте Hadoop 2.0 и его окружение со Spark и его окружением. Когда выгодно использовать один стек технологий, а когда другой?
37. Сопоставьте Spark и его окружение с SAP HANA и её окружением. Когда выгодно использовать один стек технологий, а когда другой?
38. Сопоставьте Flink и его окружение с SAP HANA и её окружением. Когда выгодно использовать один стек технологий, а когда другой?
39. Назначение и функции Modeler’a в SAP HANA. Можно ли решать какие-нибудь задачи в SAP HAMA без использования Modeler’a?
40. Какие функциональные возможности Modeler’a в SAP HANA лучше всего подходят для создания экспертных систем на базе платформы SAP HANA? Почему?
41. Сильные и слабые стороны SAP LT. Типичные варианты его применения.
42. Сильные и слабые стороны SAP Data Services (DS). Типичные варианты его применения.
43. Сильные и слабые стороны Smart Data Access. Типичные варианты его применения.
44. Сильные и слабые стороны SAP HANA Direct Extractor Connection (DXC). Типичные варианты его применения.
45. Предложите варианты загрузки данных в SAP HANA, не описанные в курсе HA350.
46. Какой проект пытается преодолеть недостатки Spark и Flink и совместить их сильные стороны. В чём заключается суть этого проекта?
47. Чем Kafka отличается от традиционных систем организации и ведения очередей?
48. Как строится эшелонированная система агентов Flume? Зачем? Какие задачи она решает? Как она отображается на структуре серверов?
49. Сопоставьте Storm и Spark Streaming. Приведите их сильные и слабые стороны.
50. Сопоставьте Storm и Samza. Приведите их сильные и слабые стороны.
51. Сопоставьте Storm и Flink. Приведите их сильные и слабые стороны.
52. Сопоставьте Spark Streaming и Samza. Приведите их сильные и слабые стороны.
53. Сопоставьте Spark и Flink. Приведите их сильные и слабые стороны.
54. Сопоставьте Samza и Flink. Приведите их сильные и слабые стороны.
55. Сопоставьте Spark Streaming и Trident. Приведите их сильные и слабые стороны.
56. Какие существуют возможности для формирования декларативных запросов, определяющих обобщённые данные, которые необходимо получить в ходе обработки потока данных? Где и каким образом это можно реализовать?
57. Основные модели данных NoSQL СУБД. Применение каждой из моделей.

58. Типы NoSQL СУБД. Применение каждого из типов.
59. Чем обусловлена востребованность NoSQL СУБД?
60. Как блокируются ресурсы средствами Zookeeper? Как можно сделать этими средствами барьер?
61. Чем отличается модель данных Zookeeper от большинства других СУБД класса Key-Value?
62. Чем отличается организация работы кластера Zookeeper от большинства других СУБД класса Key-Value?
63. Сопоставьте возможности нескольких суперстолбцовых NoSQL СУБД.
64. Как организовано взаимодействие HBase и Zookeeper?
65. Сопоставьте возможности суперстолбцовых NoSQL СУБД, традиционных SQL СУБД и Big Data SQL СУБД.
66. Особенности архитектуры и организации данных СУБД Accumulo
67. Особенности архитектуры и организации данных СУБД Cassandra.
68. Основные отличия между CQL и SQL.
69. Сопоставьте возможности нескольких документарных NoSQL СУБД.
70. Возможности поиска и манипулирования данными MongoDB.
71. Основные отличия между N1QL и SQL.
72. Сопоставьте возможности нескольких NoSQL СУБД класса Key-Value.
73. Особенности архитектуры и модели данных NoSQL СУБД Redis.
74. Возможности поиска и манипулирования данными СУБД Redis.
75. Особенности архитектуры и организации данных СУБД Kudu. Чем обусловлена более высокая скорость обработки данных в СУБД Kudu?
76. Каковы внутренние отличия SQL и NoSQL СУБД? Какие проблемы не могут быть решены средствами NoSQL СУБД и заставляют использовать SQL СУБД?
77. Какие элементы HiveQL обеспечивают более высокую скорость работы с большими данными по сравнению со стандартным SQL?
78. Почему Impala долгое время опережала по скорости работы Hive? Каким образом команда Hive пытается догнать Impala?
79. Сопоставьте возможности Phoenix и Drill. Приведите их сильные и слабые стороны.
80. В чём состоят преимущества и недостатки подхода, выбранного командами Phoenix, Drill и других аналогов к созданию SQL СУБД?
81. Сопоставьте возможности Impala и HAWQ. Приведите их сильные и слабые стороны. Когда целесообразнее использовать одну, а когда другую?
82. Чем отличаются подходы в создании и развитии SQL СУБД Impala и Vertica? Как СУБД Impala будет развиваться, чтобы достигнуть такого же состояния, что и Vertica?
83. В чём состоят преимущества и недостатки Spark SQL?
84. Чем отличаются подходы в создании и развитии SQL СУБД HP Vertica и IBM Big SQL?
85. Сопоставьте возможности SQL СУБД в составе платформ SAP HANA и Spark. Приведите их сильные и слабые стороны.
86. В чём состоят преимущества и недостатки SAP HANA Vora?
87. Почему SAP потребовалось вывести на рынок две SQL СУБД для работы с большими данными – в составе платформы SAP HANA и SAP HANA Vora? Как они могут использоваться вместе и по отдельности?
88. Сопоставьте возможности SAP HANA Vora и IBM Big SQL. Приведите их сильные и слабые стороны.
89. Сопоставьте возможности SQL СУБД, присутствующих в составе платформ Big Data (Spark, SAP HANA, Flink и др.), и возможности независимых Big Data SQL СУБД (HP Vertica, IBM Big SQL и др.). Когда целесообразнее использовать первые, а когда вторые?
90. В чём состоят три революции, имевшие место в области СУБД?

91. В чём заключаются сходства и различия процессов развития СУБД перед первой и перед второй революцией в этой области, а также в ходе каждой из них?
92. В чём заключаются сходства и различия процессов развития СУБД перед первой и перед третьей революцией в этой области, а также в ходе каждой из них?
93. В чём состоят проблемы, нерешённые в ходе третьей революции в области СУБД?
94. Чем обусловлен взрывной рост интереса в последние 2-3 года к графовым СУБД?
95. Сопоставьте возможности СУБД Neo4j, OrientDB и Titan. Приведите их сильные и слабые стороны.
96. В чём состоят различия между графовыми СУБД и Graph Engines?
97. Как соотносятся между собой между собой Giraph и Pregel?
98. В чём заключается новизна решений Surfer, GBASE, Signal/Collect и Mizan?
99. В чём состоят отличия PEGASUS от Pregel, её клонов и потомков?
100. Как проходила эволюция продуктов GraphLab? В чём новизна решений, реализуемых на каждом этапе?
101. Преимущества и недостатки Apache Spark GraphX по сравнению с независимо развиваемыми Graph Engines?
102. Каким образом Spark GraphX может использоваться для графов, хранимых в графовых СУБД?
103. Что необходимо добавить в графовую СУБД, чтобы она начала выполнять функции Graph Engine? Что необходимо добавить в Graph Engine, чтобы он начал выполнять функции графовой СУБД?
104. Сопоставьте современное состояние продуктов, сочетающих в себе возможности графовых СУБД и Graph Engines, с отдельно развиваемыми графовыми СУБД и отдельно развиваемыми Graph Engines.
105. Сопоставьте между собой SAP HANA Graph Engine и Microsoft Graph Engine, покажите их сильные и слабые стороны по отношению друг к другу.
106. Основные принципы работы и общие черты Search Engines.
107. Какую роль играют краулеры в работе Search Engines? Могут ли Search Engines работать без краулеров?
108. Общий алгоритм работы Nutch.
109. Сопоставьте возможности Solr и Elasticsearch. Приведите их сильные и слабые стороны.
110. Роль Search Engines в корпоративных системах масштаба предприятия.

Макет экзаменационного билета

Кафедра «Прикладной информатики и математики»
по дисциплине «Технологии обработки больших данных»

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Забайкальский государственный университет»

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 1
по курсу «Технологии обработки больших данных»

09.04.01 Информатика и вычислительная техника
семестр 4

1. Какие схемы обработки данных можно реализовать в YARN, которые нельзя реализовать в Hadoop 1.0? Приведите примеры таких схем. Опишите механизмы YARN, делающие возможным их реализацию.
2. Какой проект пытается преодолеть недостатки Spark и Flink и совместить их сильные стороны. В чём заключается суть этого проекта?
3. Какие элементы HiveQL обеспечивают более высокую скорость работы с большими данными по сравнению со стандартным SQL?

СОСТАВИЛ:
Доцент кафедры ПИМ
Н. А. Абдеева _____
«___» _____201__ г.

УТВЕРЖДАЮ:
Зав. кафедрой ПИМ
И.П. Глазырина_____
«__» _____ 201__ г.

Составитель _____Абдеева Н. .А.
(подпись)

4.2. Процедуры оценивания знаний, умений, навыков, формы и организация текущего контроля успеваемости и промежуточной аттестации обучающихся

Текущий контроль успеваемости и промежуточная аттестации студентов в университете ведется в соответствии с «Положением о текущем контроле успеваемости и промежуточной аттестации студентов МГТУ им. Н.Э. Баумана».

Текущий контроль успеваемости

Дисциплина в семестре делится на три модуля учебной дисциплины, включая экзамен. Каждый модуль учебной дисциплины включает в себя изучение законченного раздела, части дисциплины.

Основными видами контроля знаний, умений и навыков в течение каждого модуля учебной дисциплины являются защиты лабораторных работ, домашние задания, рубежные контроли.

Текущий контроль по модулю учебной дисциплины осуществляется по графику учебного процесса. Сроки контрольных мероприятий (КМ) и сроки подведения итогов по модулям учебной дисциплины отображаются в рабочих учебных планах на семестр (отрезках). Студент должен выполнить все контрольные мероприятия, предусмотренные в модуле учебной дисциплины к указанному сроку, после чего преподаватель проставляет балльные оценки, набранные студентом по результатам текущего контроля модуля учебной дисциплины в ЭУ.

Контрольное мероприятие считается выполненным, если за него студент получил оценку в баллах, не ниже минимальной оценки, установленной программой дисциплины по данному мероприятию.

Студенты, не сдавшие контрольное мероприятие в установленный срок, продолжают работать над ним в соответствие с порядком, принятым кафедрой.

Промежуточная аттестация

Формой промежуточной аттестации по дисциплине является экзамен.

Экзамен

Экзамен проводится в устной форме по билетам, составленным из вопросов, внесенных в программу дисциплины. На устном экзамене все дополнительные вопросы должны фиксироваться на экзаменационных листах. Учитываются рекомендации ГУИМЦ.

На экзамен выделяется 30 баллов из 100. Экзамен считается сданным, если за него студент получил в сумме не менее 18 баллов. Студент, получивший меньший балл признаётся не прошедшим промежуточную аттестацию по данной учебной дисциплине и в зачётной ведомости ему проставляется оценка «неудовлетворительно».

Оценивание дисциплины ведется в соответствии с Положением о текущем контроле успеваемости и промежуточной аттестации студентов МГТУ им. Н.Э. Баумана.

Методика оценки по рейтингу

Студент, выполнивший все предусмотренные учебным планом задания и сдавший все контрольные мероприятия, получает итоговую оценку по дисциплине за семестр в соответствии со шкалой:

Рейтинг	Оценка на экзамене
85 – 100	отлично
71 – 84	хорошо
60 – 70	удовлетворительно
0 – 59	неудовлетворительно

Рейтинг студента по дисциплине за семестр определяется как сумма баллов, полученных им за все модули учебной дисциплины, и баллов за промежуточную аттестацию. Максимальное количество баллов за дисциплину в семестре устанавливается равным 100.

ЛИСТ ВНЕСЕННЫХ ИЗМЕНЕНИЙ И ДОПОЛНЕНИЙ

<table>
<tr><td colspan="2">Номер изменения, дата внесения изменения, номер страницы для внесения изменений</td></tr>
<tr><td>БЫЛО:</td><td>СТАЛО:</td></tr>
<tr><td colspan="2">Основание:</td></tr>
<tr><td colspan="2">Подпись лица, ответственного за внесение изменений</td></tr>
</table>