**Санкт-Петербургский государственный университет**

**Р А Б О Ч А Я П Р О Г Р А М М А**

**УЧЕБНОЙ ДИСЦИПЛИНЫ**

Анализ данных

Data Analysis

**Язык(и) обучения**

русский

Трудоемкость в зачетных единицах: 4

Регистрационный номер рабочей программы: 057543

Санкт-Петербург

2020

**Раздел 1. Характеристики учебных занятий**

1. **Цели и задачи учебных занятий**

Сообщение сведений об анализе данных в объеме, необходимом для общего развития и изучения смежных дисциплин физико-математического цикла. Усвоение основных идей, понятий и фактов теории анализа данных.

**1.2. Требования подготовленности обучающегося к освоению содержания учебных занятий (пререквизиты)**

Владение курсом «Теоретическая информатика».

**1.3. Перечень результатов обучения (learning outcomes)**

Обучающийся должен овладеть теоретическим материалом в объеме, предусмотренном программой, уметь применять полученные знания при решении теоретических и прикладных задач, на основе анализа освоенных разделов: распределенные файловые системы, модель вычислений MapReduce, потоковая обработка данных и хранилища, SQL over BigData, Workflow Engines & Scheduling, архитектура систем обработки данных. Дисциплина участвует в формировании компетенций обучающихся по образовательной программе, установленных учебным планом для данной дисциплины.

**1.4. Перечень и объём активных и интерактивных форм учебных занятий**

Лекции 16, практические занятия 46 часов, промежуточная аттестация (зачёт, экзамен) 4 часа.

**Раздел 2. Организация, структура и содержание учебных занятий**

**2.1. Организация учебных занятий**

**2.1.1 Основной курс**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Трудоёмкость, объёмы учебной работы и наполняемость групп обучающихся | | | | | | | | | | | | | | | | | | |
| Код модуля в составе дисциплины,  практики и т.п. | Контактная работа обучающихся с преподавателем | | | | | | | | | | | | Самостоятельная работа | | | | Объём активных и интерактивных  форм учебных занятий | Трудоёмкость |
| лекции | семинары | консультации | практические  занятия | лабораторные работы | контрольные работы | коллоквиумы | текущий контроль | промежуточная  аттестация | итоговая аттестация | под руководством преподавателя | в присутствии  преподавателя | сам. раб. с использованием  методических материалов | текущий контроль (сам.раб.) | промежуточная аттестация (сам.раб.) | итоговая аттестация  (сам.раб.) |
| ОСНОВНАЯ ТРАЕКТОРИЯ | | | | | | | | | | | | | | | | | | |
| очная форма обучения | | | | | | | | | | | | | | | | | | |
| Семестр 6 | 16 |  | 2 | 46 |  |  |  |  | 4 |  |  |  | 40 |  | 36 |  | 52 | 4 |
|  | 2-50 |  | 2-50 | 2-50 |  |  |  |  | 2-50 |  |  |  | 1-1 |  | 1-1 |  |  |  |
| ИТОГО | 16 |  | 2 | 46 |  |  |  |  | 4 |  |  |  | 40 |  | 36 |  | 52 | 4 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Виды, формы и сроки текущего контроля успеваемости и промежуточной аттестации | | | | | | |
| Код модуля в составе дисциплины, практики и т.п. | Формы текущего контроля успеваемости | | Виды промежуточной аттестации | | Виды итоговой аттестации  (только для программ итоговой аттестации и дополнительных образовательных программ) | |
| Формы | Сроки | Виды | Сроки | Виды | Сроки |
| ОСНОВНАЯ ТРАЕКТОРИЯ | | | | | | |
| очная форма обучения | | | | | | |
| Семестр 6 |  |  | зачёт, по результатам работы за период обучения, экзамен, устно, традиционная форма | по графику промежуточной аттестации, по графику промежуточной аттестации |  |  |

**2.2. Структура и содержание учебных занятий**

Период обучения (модуль): **Семестр 6**

|  |  |  |  |
| --- | --- | --- | --- |
| № п/п | Наименование темы (раздела, части) | Вид учебных занятий | Количество часов |
| 1 | Распределенные файловые системы | Лекции | 2 |
| практические занятия | 4 |
| в присутствии преподавателя |  |
| по методическим материалам | 4 |
| 2 | Модель вычислений MapReduce | Лекции | 2 |
| практические занятия | 10 |
| в присутствии преподавателя |  |
| по методическим материалам | 8 |
| 3 | Потоковая обработка данных и хранилища | Лекции | 6 |
| практические занятия | 16 |
| в присутствии преподавателя |  |
| по методическим материалам | 14 |
| 4 | SQL over BigData | Лекции | 2 |
| практические занятия | 8 |
| в присутствии преподавателя |  |
| по методическим материалам | 6 |
| 5 | Workflow Engines & Scheduling | Лекции | 2 |
| практические занятия | 4 |
| в присутствии преподавателя |  |
| по методическим материалам | 4 |
| 6 | Архитектура систем обработки данных | Лекции | 2 |
| практические занятия | 4 |
| в присутствии преподавателя |  |
| по методическим материалам | 4 |
| 7 | Зачёт | промежуточная аттестация (ауд) | 2 |
| 8 | Экзамен | промежуточная аттестация (ауд) | 2 |
| промежуточная аттестация (с.р.) | 36 |

**Раздел 1:** Распределенные файловые системы

1. Распределенные файловые системы: об устройстве классических файловых систем и способах масштабирования до поддержки файлов на много ТБ; GFS, HDFS, почанковое хранение, иммутабельность и запрет случайных изменений, предлагаемый API распределенной файловой системы.

2. Архитектура распределенных файловых систем: мастер-сервер, ноды, пайплайн чтения и записи, локальные и нелокальные чтения; отказоустойчивость по выпадению машин (реплиакция, erasure); отказоустойчивость мастер-сервера (hot standby, shared journal, multimaster).

**Раздел 2:** Модель вычислений MapReduce

1. Предпосылки к возникновению MapReduce: модельные задачи, предшественники, commodity hardware; коммуникация как «узкое горлышко» распределенных систем обработки данных, перенос вычислений к данным; парадигмы функционального программирования.

2. Устройство MapReduce: модель вычислений; Map, Shuffle и Reduce фазы; описание пользовательских вычислений, job, task, разбиение операций на задачи; планирование задач, честный планировщик, локальность данных, stragglers.

3. Hadoop MapReduce API: Нарушение функциональных парадигм

4. Расширения модели: comparator, partitioner, combiner, зачем они нужны и когда используются.

5. Часто применяемые техники в обработке данных: map-side join, reduce-side join; salting; способы тюнинга MapReduce; способы семплирования данных; итеративные задачи.

6. Недостатки MapReduce: сostly disk spill, write barrier, job launch overhead; перекосы в данных и перекосы в планировании; от MR к DAG-ам вычислений, и почему это удобнее.

7. Spark: понятие RDD и Source RDD; computed RDD, lineage, узкие и широкие зависимости, естественная отказоустойчивость; MR over Spark, Pregel over Spark; кеширование RDD, итеративные вычисления.

**Раздел 3:** Потоковая обработка данных и хранилища

1. Kafka как «хранилище» для потоковых вычислений: модель данных, topic, partitions (as a unit of parallelism); модель отказоустойчивости, ISRs, репликация; продьюсеры и консьюмеры, стратегии партицирования и группировка консьюмеров; чтение данных at least once, обеспечение транзакциональности через durability & replay.

2. Модель вычислений Spark Streaming (Discretized Streams): аккумулирование батча и добавление его к RDD; пересчет узких и широких зависимостей.

3. Сохраняемое состояние в потоковых вычислениях.

4. Совмещение потоковой и пачковой обработки данных.

5. Модель данных BigTable/HBase (понятие строки, ключа, лексикографического порядка, колонки, семейства колонок, версии).

6. Партицирование данных, регионы (таблеты) таблицы, регион-сервера.

7. Memory Store -- in-memory append-only хранилище данных.

8. Процесс слияния версий.

9. Процесс компактификации данных.

10. Операции точечного и диапазонного чтений, операция записи.

11. Модель отказоустойчивости (WAL+Replay, синхронная репликация).

12. Примеры дизайна схемы таблицы и правильный выбора ключа.

13. Модель данных Dynamo/Cassandra.

14. Хеш-партицирование данных, consistent hashing, eventual consistency, read quorum, write quorum.

15. Антиэнтропийные техники: Hinted Handoff; Merkle Trees и их использование при синхронизации реплик.

16. ElasticSearch как хранилище данных: инвертированный индекс; шардирование и реплиакция поискового индекса; подсчет агрегированных статистик по данным; перколяция.

**Раздел 4:** SQL over BigData

1. Элементы реляционной алгебры (понятия кортежа, отношения (таблицы)).

2. Декларативное описание преобразований данных (понятие оператора, примеры).

3. SQL как синтак для описания преобразований данных.

4. Логический и физический план исполнения запроса.

5. Оптимизация плана: стоимостная модель; преобразования плана (predicate pushdown, join reordering, partial grouping).

6. Физический план исполнения SQL-запросов в MR-подобных системах (на примере Hive).

7. Физический план исполнения SQL-запросов в MPP системах: операторы SEND и EXCHANGE; преобразование и разделение дерева исполнения; роль локальности данных.

8. Физический план исполнения SQL-запросов в Spark (на примере Shark/Spark SQL): связь между RDD и операторами SQL; способы оптимизации времени работы.

**Раздел 5:** Workflow Engines & Scheduling

1. Организация сложных, многоступенчатых расчетов на кластере o WF для описания, исполнения и мониторинга процессов. Примеры: Oozie и Luigi.

2. Планирование кластерных ресурсов в многопользовательском окружении: FIFO-планирование, честное планирование; механизмы обеспечения ресурсных гарантий для критичных процессов.

**Раздел 6:** Архитектура систем обработки данных

1. Трейдоффы в batch и real-time обработке данных.

2. Часто используемые техники экономии сложности вычисления (вероятностные структуры данных).

3. Лямбда-архитектура.

**Раздел 3. Обеспечение учебных занятий**

**3.1. Методическое обеспечение**

**3.1.1 Методические указания по освоению дисциплины**

Посещение лекций и практических занятий.

**3.1.2 Методическое обеспечение самостоятельной работы**

Основная и дополнительная литература.

**3.1.3 Методика проведения текущего контроля успеваемости и промежуточной аттестации и критерии оценивания**

*Методика проведения зачёта*

Зачет проводится в устной форме. Для получения зачета необходимо решить 60% задач, предлагаемых в течение семестра. В случае, если к моменту проведения зачета студент решил меньшее количество задач, на зачете ему предлагаются задачи аналогичные по тематике и сложности. Задачи даются в форме домашних заданий с устной сдачей («листочки»), письменных домашних заданий и контрольных. Темы задач фиксированы, количество и форма выдачи остается на усмотрение преподавателя практических занятий. Возможна выдача задач повышенной сложности, решение которых засчитывается в качестве индивидуальных достижений студента (при подаче заявок на именные стипендии, конкурсы и т.п.); сдача таких заданий проводится в устной форме.

*Методика проведения экзамена*

Экзамен проводится в устной форме. Билет состоит из двух вопросов. Время подготовки ответа на вопросы билета составляет 60 минут.

Использование конспектов и учебников, а также электронных устройств хранения, обработки или передачи информации при подготовке и ответе на вопросы экзамена категорически запрещено. В случае обнаружения факта использования недозволенных материалов (устройств) составляется акт и студент удаляется с экзамена. После ответа на вопросы билета преподаватель задает несколько дополнительных вопросов, на основании оценки ответов на которые итоговая оценка по предмету может быть повышена или понижена.

*Критерии выставления оценок*

Оценка «отлично» ставится за полностью раскрытый теоретический материал и правильные ответы на дополнительные вопросы преподавателя. В болонской шкале оценка может быть скорректирована в ту или иную сторону с учетом малозначительных погрешностей изложения или, напротив, углубленного изложения материала.

Оценка «хорошо» ставится за изложенный теоретический материал билета (возможно с помощью наводящих подсказок преподавателя).

Оценка «удовлетворительно» ставится за знание основных вопросов по каждой теме.

Оценка «неудовлетворительно» выставляется, если не выполняются условия для получения оценок «отлично», «хорошо» и «удовлетворительно».

Соответствие оценки СПбГУ и оценки ECTS (Европейской системы переноса и накопления зачётных единиц):

|  |  |  |  |
| --- | --- | --- | --- |
| Итоговый процент выполнения, % | Оценка СПбГУ при  проведении зачёта | Оценка ECTS | Оценка СПбГУ при  проведении экзамена |
| 90-100 | зачтено | A | отлично |
| 80-89 | зачтено | B | хорошо |
| 70-79 | зачтено | C | хорошо |
| 60-69 | зачтено | D | удовлетворительно |
| 50-59 | зачтено | E | удовлетворительно |
| менее 50 | не зачтено | F | неудовлетворительно |

**3.1.4 Методические материалы для проведения текущего контроля успеваемости и промежуточной аттестации (контрольно-измерительные материалы, оценочные средства)**

Период обучения (модуль): **Семестр 6**

Темы **задач**:

1. Распределенные файловые системы и их архитектура.

2. Модель вычислений MapReduce и её расширения.

3. Недостатки MapReduce. Spark.

4. Потоковая обработка данных.

5. BigTable-подобные хранилища, HBase.

6. Dynamo-подобные хранилища, Cassandra.

7. Специализированные хранилища, ElasticSearch.

8. SQL over BigData

9. Workflow Engines & Scheduling.

10. Архитектура систем обработки данных.

**Список вопросов к экзамену:**

1. Распределенные файловые системы: об устройстве классических файловых систем и способах масштабирования до поддержки файлов на много ТБ; GFS, HDFS, почанковое хранение, иммутабельность и запрет случайных изменений, предлагаемый API распределенной файловой системы.

2. Архитектура распределенных файловых систем: мастер-сервер, ноды, пайплайн чтения и записи, локальные и нелокальные чтения; отказоустойчивость по выпадению машин (реплиакция, erasure); отказоустойчивость мастер-сервера (hot standby, shared journal, multimaster).

3. Предпосылки к возникновению MapReduce: модельные задачи, предшественники, commodity hardware; коммуникация как «узкое горлышко» распределенных систем обработки данных, перенос вычислений к данным; парадигмы функционального программирования.

4. Устройство MapReduce: модель вычислений; Map, Shuffle и Reduce фазы; описание пользовательских вычислений, job, task, разбиение операций на задачи; планирование задач, честный планировщик, локальность данных, stragglers.

5. Hadoop MapReduce API: Нарушение функциональных парадигм

6. Расширения модели: comparator, partitioner, combiner, зачем они нужны и когда используются.

7. Часто применяемые техники в обработке данных: map-side join, reduce-side join; salting; способы тюнинга MapReduce; способы семплирования данных; итеративные задачи.

8. Недостатки MapReduce: сostly disk spill, write barrier, job launch overhead; перекосы в данных и перекосы в планировании; от MR к DAG-ам вычислений, и почему это удобнее.

9. Spark: понятие RDD и Source RDD; computed RDD, lineage, узкие и широкие зависимости, естественная отказоустойчивость; MR over Spark, Pregel over Spark; кеширование RDD, итеративные вычисления.

10. Kafka как «хранилище» для потоковых вычислений: модель данных, topic, partitions (as a unit of parallelism); модель отказоустойчивости, ISRs, репликация; продьюсеры и консьюмеры, стратегии партицирования и группировка консьюмеров; чтение данных at least once, обеспечение транзакциональности через durability & replay.

11. Модель вычислений Spark Streaming (Discretized Streams): аккумулирование батча и добавление его к RDD; пересчет узких и широких зависимостей.

12. Сохраняемое состояние в потоковых вычислениях.

13. Совмещение потоковой и пачковой обработки данных.

14. Модель данных BigTable/HBase (понятие строки, ключа, лексикографического порядка, колонки, семейства колонок, версии).

15. Партицирование данных, регионы (таблеты) таблицы, регион-сервера.

16. Memory Store -- in-memory append-only хранилище данных.

17. Процесс слияния версий.

18. Процесс компактификации данных.

19. Операции точечного и диапазонного чтений, операция записи.

20. Модель отказоустойчивости (WAL+Replay, синхронная репликация).

21. Примеры дизайна схемы таблицы и правильный выбора ключа.

22. Модель данных Dynamo/Cassandra.

23. Хеш-партицирование данных, consistent hashing, eventual consistency, read quorum, write quorum.

24. Антиэнтропийные техники: Hinted Handoff; Merkle Trees и их использование при синхронизации реплик.

25. ElasticSearch как хранилище данных: инвертированный индекс; шардирование и реплиакция поискового индекса; подсчет агрегированных статистик по данным; перколяция.

26. Элементы реляционной алгебры (понятия кортежа, отношения (таблицы)).

27. Декларативное описание преобразований данных (понятие оператора, примеры).

28. SQL как синтак для описания преобразований данных.

29. Логический и физический план исполнения запроса.

30. Оптимизация плана: стоимостная модель; преобразования плана (predicate pushdown, join reordering, partial grouping).

31. Физический план исполнения SQL-запросов в MR-подобных системах (на примере Hive).

32. Физический план исполнения SQL-запросов в MPP системах: операторы SEND и EXCHANGE; преобразование и разделение дерева исполнения; роль локальности данных.

33. Физический план исполнения SQL-запросов в Spark (на примере Shark/Spark SQL): связь между RDD и операторами SQL; способы оптимизации времени работы.

34. Организация сложных, многоступенчатых расчетов на кластере o WF для описания, исполнения и мониторинга процессов. Примеры: Oozie и Luigi.

35. Планирование кластерных ресурсов в многопользовательском окружении: FIFO-планирование, честное планирование; механизмы обеспечения ресурсных гарантий для критичных процессов.

36. Трейдоффы в batch и real-time обработке данных.

37. Часто используемые техники экономии сложности вычисления (вероятностные структуры данных).

38. Лямбда-архитектура.

**3.1.5 Методические материалы для оценки обучающимися содержания и качества учебного процесса**

Анкета для студентов для оценки качества преподавания курса.

Просим Вас заполнить анкету-отзыв по прочитанной дисциплине. Обобщенные данные анкет будут использованы для ее совершенствования. По каждому вопросу проставьте соответствующие оценки по шкале от 1 до 10 баллов (обведите выбранный Вами балл). В

случае необходимости впишите свои комментарии.

1. Насколько Вы удовлетворены содержанием дисциплины в

целом?

1 2 3 4 5 6 7 8 9 10

Комментарий\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2. Насколько Вы удовлетворены общим стилем преподавания?

1 2 3 4 5 6 7 8 9 10

Комментарий\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3. Как Вы оцениваете качество подготовки предложенных

методических материалов?

1 2 3 4 5 6 7 8 9 10

Комментарий\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4. Насколько Вы удовлетворены использованием

преподавателями активных методов обучения?

1 2 3 4 5 6 7 8 9 10

Комментарий\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5. Какой из модулей (разделов) дисциплины Вы считаете наиболее полезным, ценным с точки зрения дальнейшего обучения и/или

применения в последующей практической деятельности?

Комментарий\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6. Что бы Вы предложили изменить в методическом и

содержательном плане для совершенствования преподавания данной

дисциплины?

Комментарий\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

СПАСИБО!

**3.2. Кадровое обеспечение**

**3.2.1 Образование и (или) квалификация штатных преподавателей и иных лиц, допущенных к проведению учебных занятий**

К чтению лекций должны привлекаться преподаватели, имеющие ученую степень доктора или кандидата наук (в том числе степень PhD, прошедшую установленную процедуру признания и установления эквивалентности) и/или ученое звание профессора или доцента.

**3.2.2 Обеспечение учебно-вспомогательным и (или) иным персоналом**

Не требуется.

**3.3. Материально-техническое обеспечение**

**3.3.1 Характеристики аудиторий (помещений, мест) для проведения занятий**

Учебные аудитории для проведения учебных занятий, оснащенные стандартным оборудованием, используемым для обучения в СПбГУ в соответствии с требованиями материально-технического обеспечения.

**3.3.2 Характеристики аудиторного оборудования, в том числе неспециализированного компьютерного оборудования и программного обеспечения общего пользования**

Стандартное оборудование, используемое для обучения в СПбГУ. MS Windows, MS Office, Mozilla FireFox, Google Chrome, Acrobat Reader DC, WinZip, Антивирус Касперского.

**3.3.3 Характеристики специализированного оборудования**

Доступ к компьютерному кластеру высокопроизводительных вычислений.

**3.3.4 Характеристики специализированного программного обеспечения**

Не требуется.

**3.3.5 Перечень и объёмы требуемых расходных материалов**

Мел — не менее 1 куска на час лекционных занятий, фломастеры для доски, губка.

**3.4. Информационное обеспечение**

**3.4.1 Список обязательной литературы**

1. Граннеман, С. Linux. Карманный справочник. — 2-е изд. — Вильямс, 2016. — 464 с.

2. Немет, Э., Снайдер, Г., Хейн, Т.Р., и Уэйли, Б. Unix и Linux. Руководство системного администратора. — 4-е изд. — Вильямс, 2014. — 1312 с.

3. Helmke, М., Joseph, E.K., and Rey, J.A., Ballew, P., with Hill, B.M. The Official Ubuntu Book. — 8th ed. — Prentice Hall, 2014. — 368 p.

**3.4.2 Список дополнительной литературы**

1. Lublinsky, B., Smith, K.T., and Yakubovich, A. Professional Hadoop Solutions. — Wrox, 2013. — 504 p.

2. White, T. Hadoop: The Definitive Guide. — 3rd ed. — O’Reilly, 2012. — 688 p.

**3.4.3 Перечень иных информационных источников**

1. Сайт Научной библиотеки им. М. Горького СПбГУ: http://www.librarv.spbu.ru/

2. Электронный каталог Научной библиотеки им. М. Горького СПбГУ: http://www.librarv.spbu.ru/cgibin/irbis64r/cgiirbis 64.ехе?С21 COM=F&I21 DBN=IBIS&P21 DBN=IBIS

3. Перечень электронных ресурсов, находящихся в доступе СПбГУ: http://cufts.librarv.spbu.ru/CRDB/SPBGU/

4. Перечень ЭБС, на платформах которых представлены российские учебники, находящиеся в доступе СПбГУ: http://cufts.library.spbu.ru/CRDB/SPBGU/browse?name=rures&resource tvpe=8

**Раздел 4. Разработчики программы**

Гирш Эдуард Алексеевич, доктор физико-математических наук, профессор СПбГУ, hirsch@pdmi.ras.ru