АННОТАЦИЯ РАБОЧЕЙ ПРОГРАММЫ ДИСЦИПЛИНЫ
СБОР ОБРАБОТКА И ХРАНЕНИЕ РАСПРЕДЕЛЕННЫХ ДАННЫХ

Автор: Стефановский Дмитрий Владимирович
Код и наименование направления подготовки, профиля: 38.04.01 Экономика («Системы больших данных в экономике»)
Квалификация (степень) выпускника: Магистр
Форма обучения: очная

Цель освоения дисциплины:

Сформировать компетенции в сфере аналитической работы в области экономики и финансов, обработки и анализов данных.

План курса:

Тема 1. Распределенный анализ данных.

Системы мобильных агентов. Основные понятия. Стандарты многоагентных систем. Системы мобильных агентов. Система мобильных агентов JADE. Использование мобильных агентов для анализа данных. Проблемы распределенного анализа данных. Агенты-аналитики. Варианты анализа распределенных данных.

Система анализа распределенных данных. Общий подход к реализации системы. Агент для сбора информации о базе данных. Агент для сбора статистической информации о данных. Агент для решения одной задачи интеллектуального анализа данных. Агент для решения интегрированной задачи интеллектуального анализа данных.

Тема 2. Data Mining в реальном времени (Real-Time Data Mining).

Идея Data Mining в реальном времени. Адаптация системы к общей концепции. Адаптивная добыча данных. Статический Data Mining и Data Mining в реальном времени. Применение Data Mining в реальном времени.
Рекомендательные машины. Классификация рекомендательных машин. Подход на основе содержания. Совместное фильтрование. Анализ рыночной корзины и секвенциальный анализ. Усиление обучения и агенты.

Инструменты Data Mining в реальном времени

Тема 3. Извлечение знаний из Web — Web Mining

Web Mining. Проблемы анализа информации из Web. Проблемы анализа информации из Web. Этапы Web Mining. Web Mining и другие интернет-технологии. Категории Web Mining.

Методы извлечения Web-контента. Извлечение Web-контента в процессе информационного поиска. Извлечение Web-контента для формирования баз данных

Извлечение Web-структур. Представление Web-структур. Оценка важности Web-структур. Поиск Web-документов с учетом гиперссылок. Кластеризация Web-структур.

Исследование использования Web-ресурсов. Исследуемая информация. Этап препроцессинга. Этап извлечения шаблонов. Этап анализа шаблонов и их применение.

Тема 4. Программирование операций с RDD. Работа с парами ключ/значение

Основы RDD. Создание RDD. Операции с RDD. Преобразования. Действия. Отложенные вычисления. Передача функций в Spark. Применры на: Python, Scala, Java. Простые наборы RDD. Преобразование типов RDD. Сохранение (кэширование).

Создание наборов пар. Преобразования наборов пар. Агрегирование. Группировка данных. Соединения. Сортировка. Действия над наборами пар ключ/значение. Управление распределением данных. Определение объекта управления распределением RDD. Операции, получающие выгоды от наличия информации о распределении. Операции, на

которые влияет порядок распределения. Пример: PageRank. Собственные объекты управления распределением.

Тема 5. Выполнение в кластере

Архитектура среды Spark времени выполнения. Драйвер. Исполнители. Диспетчер кластера. Запуск программы. Итоги. Развертывание приложений с помощью spark-submit. Упаковка программного кода и зависимостей. Сборка приложения на Java с помощью Maven. Сборка приложения на Scala с помощью sbt. Конфликты зависимостей. Планирование приложений в Spark. Диспетчеры кластеров. Диспетчер кластера Spark Standalone. Hadoop YARN. Apache Mesos. Amazon EC2. Выбор диспетчера кластера.

Тема 6. Spark SQL

Включение Spark SQL в приложения. Использование Spark SQL в приложениях. Инициализация Spark SQL. Пример простого запроса. Наборы данных SchemaRDD. Кэширование. Загрузка и сохранение данных. Apache Hive. Parquet. JSON. Из RDD. Сервер JDBC/ODBC. Работа с программой Beeline. Долгоживущие таблицы и запросы. Функции, определяемые пользователем. Spark SQL UDF. HiveUDF. Производительность Spark SQL.

Тема 7. Spark Streaming

Архитектура и абстракция. Преобразования без сохранения состояния. Преобразования с сохранением состояния. Операции вывода. Источники исходных данных. Основные источники. Дополнительные источники. Множество источников и размеры кластера. Круглосуточная работа. Копирование в контрольных точках. Повышение отказоустойчивости драйвера. Отказоустойчивость рабочих узлов. Отказоустойчивость приемников. Гарантированная обработка. Веб-интерфейс Spark Streaming. Проблемы производительности. Интервал пакетирования и протяженность окна. Степень параллелизма. Сборка мусора и использование памяти.

Тема 8. Машинное обучение с MLlib

Системные требования. Основы машинного обучения. Пример: классификация спама. Типы данных. Векторы. Алгоритмы. Извлечение признаков. Статистики. Классификация и регрессия. Кластеризация. Коллаборативная фильтрация и рекомендации. Понижение размерности. Оценка модели. Советы и вопросы производительности. Выбор признаков. Настройка алгоритмов. Кэширование наборов RDD для повторного использования. Разреженные векторы. Степень параллелизма. Высокоуровневый API машинного обучения.

Аудиторные часы: 180

Формы текущего контроля и промежуточной аттестации: опросы, домашнее задание, контрольная работа, экзамен.

Основная литература:

1. Карау, Холден. Изучаем Spark: молниеносный анализ данных / Холден Карау, Энди Конвински, Патрик Венделл, Матей Захария. - Москва: ДМК Пресс, 2015. - 304 с.: ил.; ISBN 978-5-97060-323-9