



ПОЛИТЕХ

Санкт-Петербургский
Политехнический Университет
Петра Великого

**Санкт-Петербургский политехнический университет Петра Великого
Институт компьютерных наук и технологий
Высшая школа программной инженерии**

Системы анализа больших данных (САБД)

Введение в дисциплину

Направление: *09.04.04 – «Программная инженерия»*

Преподаватель Ковалев Артем Дмитриевич

Цель курса

- Цель изучения дисциплины «Системы анализа больших данных» направлена на:
 - формирование у обучающихся пониманий и знаний теоретических и практических аспектов и подходов к проектированию и реализации комплексных программных систем по анализу данных, а также проблем и подходов их решения, которые адресуются в системах анализа больших объемов данных.
 - подготовку квалифицированных выпускников, умеющих эффективно и качественно разрабатывать и внедрять программные комплексы и инструментальные средства по анализу и работе с информацией.
 - выработку навыков самостоятельного исследования и изучения технологий, систем, программных комплексов, архитектур, программных особенностей API в САБД
 - формирование умений реализации современных подходов, используемых при проектировании систем обработки больших данных

Структура курса

Курс состоит из двух частей:

- теоретическая часть

подготовка материала и выступление перед аудиторией по выбранной тематике

- практическая часть

реализация современных подходов проектирования программного обеспечения, используемых в системах обработки больших данных

Варианты тем для выступления (1)

- Системы анализа больших данных:

- ! IBM Watson
- ! Виртуальный помощник IPSoft Amelia
- ! Когнитивные системы помощи клиентам (Чат боты, поддержка у Мегафон, и т.д.)
- Semantext
- Dell EMC Analytic Insights Module
- Windows Azure HDInsight
- Microsoft Azure Machine Learning
- Pentaho Data Integration
- Teradata Aster Analytics
- SAP BusinessObjects Predictive Analytics
- Oracle Big Data Preparation
- другие

Используйте Google,
сайты поставщиков,
книги, статьи и форумы

Варианты тем для выступления (2)

- Базы данных:
 - Apache Hive
 - Cloudera Impala
 - Apache Presto
 - Apache Drill
 - Apache Cassandra
 - Redis
 - EMC Greenplum
 - другие

Варианты тем для выступления (3)

- Аналитические платформы:
 - RapidMiner
 - IBM SPSS Modeler
 - KNIME
 - Qlik Analytics Platform
 - STATISTICA Data Miner
 - Informatica Intelligent Data Platform
 - World Programming System
 - Deductor
 - SAS Enterprise Miner
 - другие

Варианты тем для выступления (4)

- **Фреймворки:**
 - Elasticsearch
 - Kibana
 - Apache Flink
 - Apache ZooKeeper
 - Apache Mesos
 - Apache Flume
 - другие
- **Аварийное восстановление ("disaster recovery") программных систем после сбоев**
 - **!** Обзор существующих подходов и методов
 - **!** Существующие программные системы, сравнительный анализ реализаций и ограничений

Варианты тем для выступления (5)

- Способы повышения безопасности работы с данными:
 - Способы обфускирования и обезличивания информации
 - Применение шифрования данных стандартными библиотеками: BouncyCastle, SafeNet Keysecure Gemalto и SunJCE
 - Использование безопасных соединений по протоколам HTTPS с использованием ключей шифрования для SSL/TLS (Two-way TLS)
 - Локализация распределенных программных систем анализа в выделенной, изолированной локальной сети. Подход применения Gateway для выхода из изолированной сети во внешний мир

Варианты тем для выступления (6)

- Экономические трудности применения облачных и кластерных систем анализа
 - Сравнительный анализ способов развертывания программных систем по экономическим показателям и функциональным возможностям на стороне заказчика "On-premises", в удаленном облаке, у сторонней организации предоставляющей необходимые сервисы и вычислительные мощности

План выступления

- 20 минут на одно выступление
- Обзор системы/технологии/инструмента
 - назначение
 - возможности
 - ограничения
- Задачи, которые можно решить
- Программная архитектура и основные модули системы
- Пример использования
- Обзор API

Требования к выступлениям

- Предварительная запись на выступление
 - староста делаает табличку со списком групп и разлиновкой по неделям занятий в Google таблицах
 - студенты бронируют тему и заносят себя в определенный день доклада
 - в один день по 3-4 выступления
 - ppt-версия презентации выкладывается в группу VK

Запись на выступления

ФИО/Дата	09.09	16.09	23.09	30.09	...
Вася (гр №)	+				
Петя (гр №)		+			
Коля (гр №)	+				
Оля (гр №)		+			

ФИО	Тема
Вася (гр №)	Способы обфускирования и обезличивания информации
Петя (гр №)	...
Коля (гр №)	...
Оля (гр №)	...

Практические задачи

1. Реализация программного средства для обфускировки и де-обфускирования данных
2. Создание демонстрационной программы и тестовых сценариев по шифрованию данных стандартной библиотекой BouncyCastle
3. Проектирование и реализация клиент-серверного приложения, взаимодействующего по HTTPS протоколу с использованием ключей шифрования для SSL/TLS (Two-way TLS)
4. Создание маршрутизатора для клиент-серверного приложения, работающего через Gateway по средствам библиотеки Netflix Zuul.

Реализация практических задач

- Без отчетов
- Ссылку на репозиторий GitHub в обсуждение VK
- Демонстрация работы

Правила оценки успеваемости

Активность	Баллы за 1 ед.	Мах.баллов
Посещение занятий	2	30
Выступление с докладом	30	30
Практические задачи	10	40

Набрано баллов	Оценка
до 40	Неудовлетворительно
40 - 60	Удовлетворительно
60 - 80	Хорошо
80 - 100	Отлично

Спасибо за внимание!

Вопросы?