Системы хранения и обработки потоков больших данных

Синхронизированный план лекций, лабораторных работ и семинаров для 2 курса магистратуры ИУ-6 на 2020/2021 уч. год

Нед.	Лекции	Лабораторные/семинарские работы (4 часа)
1	Введение в потоковую обработку данных.	
2	Основы программирования процессов обработки. Организация потоков данных. Модели программирования. Операции с окном и триггеры.	Создание процесса обработки данных с помощью Apache Storm / Apache Samza.
3	Хранилища потоковых данных. HDFS, Kafka, Pravega, Ignite Persistence, ...	Создание процесса обработки с помощью Apache Flink.
4	Методы обеспечения устойчивости ко сбоям. Протоколы синхронизации. Хранение состояния. RocksDB, специализированные харанилища Оперативный доступ к данным.	Хранилища потоковых данных. HDFS, Kafka, Pravega, Ignite Persistence.
5	Задачи оптимизации плана обработки данных. Исследование продуктов к способности оптимизации Балансирование нагрузки в распределенной системе. Планирование аппаратных ресурсов Визуалиазция топологии процесса обработки данных Задачи машинного обучения с потоковыми данными Оценка рабочих характеристик программно-аппаратного комплекса. Мониторинг распределенной системы. Обратное давление. Отладка и журналирование	Исследование бенчмарков средств потоковой обработки Измерения параметров обслуживания потока: времени обслуживания заявки, задержки начала обслуживания, оценивание обратного давление.
6	Решение типовых задач. Интеграционные продукты (например Apache Beam)	

ДЗ 1 — решение типовой задачи обработки потока данных при помощи потокового процессора — Storm, Samza, Flink, Spark, Apex...
ДЗ 2 — создание стенда для оценки производительности работы потокового процессора в связке со слоем накопления и хранения информации