

## **ОБУЧЕНИЕ ТЕХНОЛОГИЯМ ОБРАБОТКИ И АНАЛИЗА БОЛЬШИХ ДАННЫХ В СРЕДЕ ВИРТУАЛЬНОЙ КОМПЬЮТЕРНОЙ ЛАБОРАТОРИИ В ГОСУДАРСТВЕННОМ УНИВЕРСИТЕТЕ «ДУБНА»**

**Белов М.А.<sup>а</sup>, Гришко С.И.<sup>б</sup>, Токарева Н.А.<sup>с</sup>**

Государственное бюджетное образовательное учреждение высшего образования  
Московской области «Университет «Дубна»;  
РФ, 141980, Московская область, г. Дубна, ул. Университетская, д. 19, к. 1-312;  
+ 7 (496) 2166010, <sup>а</sup>belov@uni-dubna.ru; <sup>б</sup>si.10@yandex.ru; <sup>с</sup>tokareva@uni-dubna.ru

На сегодняшний день наблюдается устойчивый рост спроса на специалистов, способных быстро создавать программно-технологические решения в области обработки больших данных, поскольку всё больше количество организаций испытывает потребность анализа данных объемом до десятков петабайт, с целью быстрой обработки статистических показателей – абсолютных, относительных и средних величин, вариаций и статистических индексов, а также выявления причинно-следственных связей и закономерностей, оценивать эффективность и возможности экономических и социальных явлений.

Одной из основополагающих технологий обработки больших данных является Hadoop — продукт фонда Apache Software Foundation, состоящий из свободно распространяемых утилит, библиотек и фреймворков с открытым кодом. Вокруг Hadoop образовалась целая экосистема из связанных проектов и технологий, что позволяет быстро развернуть кластер и обеспечить его продуктивную эксплуатацию. Обучение этим современным технологиям в институте системного анализа и управления государственного университета «Дубна» проводится на базе «Виртуальной компьютерной лаборатории» (ВКЛ) — инновационного инструмента, разработанного в институте, и представляющего собой комплекс программно-аппаратных средств, основанных на технологиях виртуализации, контейнеризации и управления знаниями. Использование ВКЛ позволяет гибко, по запросу, предоставлять и использовать вычислительные ресурсы, эффективно реализовывать и использовать многокомпонентные системы, выполнять ресурсоемкие вычислительные аналитические расчеты в рамках аудиторной и внеаудиторной работы.

В докладе обсуждается выбор основных компонентов программно-технологического решения Hadoop для внедрения в учебный процесс. Для выполнения распределенных вычислений в рамках парадигмы MapReduce с применением языка программирования Java используется программный фреймворк MapReduce. Последний выбран потому, что студенты уже достаточно знакомы с принципами объектно-ориентированного программирования и языком программирования Java и могут сосредоточиться на изучении основных аналитических паттернов и конструкций для реализации базового обработчика, который на каждом вычислительном узле кластера обеспечивает преобразование исходных пар «ключ — значение» в промежуточный набор пар «ключ — значение» в классе, реализующим интерфейс Mapper и названный по функции высшего порядка Map, и обработчика который преобразует промежуточный набор пар в итоговый сокращенный набор — свёртку, класс, реализующий интерфейс Reducer, с возможностью записи результата в реляционную базу данных.