

УДК 004.41 (075.8)

ББК 22.18 я73

Б 17

Печатается по решению
редакционно-издательского совета
Северо-Кавказского федерального
университета

Б 17 Базы данных в высокопроизводительных информационных системах: учебное пособие / авт.-сост. Е. И. Николаев. – Ставрополь: Изд-во СКФУ, 2016. – 163 с.

Пособие составлено в соответствии с требованиями Федерального государственного образовательного стандарта, программой и учебным планом дисциплины. Содержит теоретические аспекты проектирования и разработки приложений для высокопроизводительных вычислительных систем, в основе функционирования которых лежат базы данных. Основные технологии, рассматриваемые в пособии, – Hadoop, NoSQL.

Предназначено для студентов направления подготовки 09.04.02 – Информационные системы и технологии, обладающих теоретическими знаниями в области проектирования высокопроизводительных приложений и практическими навыками программирования (предпочтительно языки Java, SQL, Python).

УДК 004.41 (075.8)

ББК 22.18 я73

Автор-составитель

канд. техн. наук, доцент Е. И. Николаев

Рецензенты:

д-р физ.-мат. наук, профессор В. И. Дроздова,

д-р техн. наук, профессор А. В. Маликов

© ФГАОУ ВО Северо-Кавказский
федеральный университет, 2016

ПРЕДИСЛОВИЕ

Современные технологии обработки и хранения информации не ограничиваются только реляционными данными. Непрерывное увеличение объемов обрабатываемых данных приводит к поиску новых программных и аппаратных решений для обеспечения высоких характеристик хранилищ информации. Одним из таких подходов является применение специализированных систем управления базами данных (СУБД), существенно отличающихся от ранее используемых – это нереляционные базы данных. СУБД, разрабатываемые в рамках данного подхода, получили название NoSQL СУБД. Несмотря на принципиальные отличия от ранее применяемых, представители технологии NoSQL позволяют при грамотном применении обеспечивать существенные преимущества по сравнению с реляционными СУБД. Подход NoSQL затронул программную часть информационных систем; а также существенно пересмотрены принципы построения центров обработки данных для использования в рамках NoSQL.

В рамках развития высокопроизводительных систем обработки данных появились сложные технологии, которые не относятся к СУБД, например, технология Hadoop. Данный фреймворк представляет собой целый стек технологий, обеспечивающих слой виртуализации и сервисов для высокопроизводительных программных комплексов. Hadoop рассчитан на применение в распределенной гетерогенной среде, данная технология не предполагает применения какого-либо одного языка программирования или СУБД – это именно уровень в многослойной информационной системе.

Потребность в высокопроизводительных, распределенных и масштабируемых СУБД, вызванная объективными причинами техногенного характера, приводит к неуклонному росту востребованности специалистов по работе с данными, аналитиков в области Big Data на рынках труда практически всех стран. Большинство современных университетов предоставляют возможность подготовки специалистов по направлениям Big Data и NoSQL.

Таким образом, при проектировании высокопроизводительных информационных систем на основе баз данных необходимо учитывать тенденции и современное состояние программных и аппаратных средств обработки данных.

В пособии рассматривается комплекс вопросов, направленных на всестороннее изучение процессов, протекающих в области Data Science. Изложение материала пособия построено по принципу «от теории к практике». Первые разделы отражают исторические аспекты NoSQL-подхода; теоретические подходы, лежащие в основе науки о данных. Последние посвящены изложению методов работы с высокопроизводительными системами обработки данных уровня предприятия.

Пособие позволяет учащемуся получить исчерпывающие теоретические представления о таких технологиях, как Hadoop, NoSQL, а также обеспечивает магистра достаточной информацией для самостоятельных исследований в области Big Data.

Материал, размещенный в пособии, способствует формированию следующих компетенций:

- 1) ОК-3 – умение свободно пользоваться русским и иностранным языками как средством делового общения;
- 2) ОПК-4 – владением по крайней мере одним из иностранных языков на уровне социального и профессионального общения, способностью применять специальную лексику и профессиональную терминологию языка.

СОДЕРЖАНИЕ

Предисловие	3
Раздел 1. Основы BIG DATA	5
Раздел 2. Модели и концепции BIG DATA	43
Раздел 3. Основы NOSQL	59
Раздел 4. Фреймворк распределенного программирования ...	85
Раздел 5. YARN	102
Раздел 6. Паттерны MAPREDUCE и BIG DATA	123
Раздел 7. MONGODB	137
Заключение	160
Литература	161