ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ И РАДИОЭЛЕКТРОНИКИ

И.В. Бойченко

«Высокопроизводительные распределенные системы»

Методические указания к практическим занятиям и по
организации самостоятельной работы магистрантов,
обучающихся по направлению «Программная инженерия»

2016

Оглавление

Введение....3
1 Технологии Big Data....5
2 Практические занятия....7
2.1 Занятие №1 Параллельное программирование в Java....7
2.2 Занятие №2 Технология Hadoop и Map/Reduce....7
2.3 Занятие №3 Технология Apache Spark + Java....8
2.4 Занятие №4 Технология Apache Spark + R....8
2.5 Занятие №5 Технологии Apache Hive, Hbase....9
3 Самостоятельная работа....10
3.1 Темы на самостоятельное изучение....10
3.2 Темы рефератов....10
Литература....11

Введение

Цель изучения дисциплины состоит в формировании знаний умений и навыков в области разработки и эксплуатации программного обеспечения современных высокопроизводительных распределенных систем. В данном курсе рассматриваются программные технологии построения масштабируемых многомашинных инфомационно-вычислительных систем, обеспечивающих параллельную обработку сверхбольших массивов данных. За рубежом совокупность таких технологий обозначается термином Big Data (англ. - большие данные).

Рассматриваются также типовые методы и алгоритмы параллельной обработки сверхбольших массивов данных с использование стека технологий Big Data.

Задачи изучения дисциплины:

1) ознакомление с теоретическими основами организации параллельной распределенной обработки данных на программном уровне;

2) получение опыта практической работы с современными программными инструментами для параллельной распределенной обработки данных.

Дисциплина **«Высокопроизводительные распределенные системы»** относится к обязательным дисциплинам вариативной части структуры основных профессиональных образовательных программ (ОПОП). Для успешного освоения данной дисциплины необходимо и достаточно знаний и умений, приобретенных студентами при изучении на предыдущем уровне образования таких дисциплин, как «Объектно-ориентированный анализ и программирование», «Вычислительные системы, сети и телекоммуникации», «Операционные системы».

Дисциплина является базовой при проведении научно-исследовательской работы магистра, прохождении научно-исследовательской практики, подготовке магистерской диссертации.

В результате изучения дисциплины студент должен **знать:**

1. теоретические основы организации распределенных вычислений;
2. состав и принципы построения ПО параллельных распределенных вычислений;
3. методы измерения производительности вычислительных систем;

уметь:

4. реализовывать параллельные алгоритмы обработки данных на высокоуровневых языках программирования с использованием библиотек;

5. устанавливать и настраивать окружение распределенных вычислений с использованием современных программных продуктов;

владеть:

6. средствами выполнения и отладки прикладного ПО для распределенных систем;
7. средствами профилирования и измерения производительности при решении задач на распределенных вычислительных системах.

1 Технологии Big Data

Big Data (англ. - большие данные) – объединяющее название стека технологий ориентированного на обработку данных, характеризующихся тремя критериями («3V»)[1]: объем (Volume), скорость (Velocity) и вариативность (Variety).

Независимо от реализации, в основу технологий Big Data положены два основных принципа:

1. принцип распределенного хранения данных;
2. принцип распределенной обработки, с учетом локальности данных;

Распределенное хранение решает проблему большого объема данных, позволяя организовывать хранилище из произвольного числа отдельных простых носителей, как правило, обычных жестких дисков. Хранение может быть организовано с разной степенью избыточности, обеспечивая устойчивость к сбоям отдельных носителей.

Распределенная обработка с учетом локальности данных означает, что программа обработки доставляется на вычислитель, находящийся как можно ближе к обрабатываемым данным. Это принципиально отличается от традиционного подхода, когда вычислительные мощности и подсистема хранения разделены, и данные должны быть доставлены на вычислитель.

Таким образом, технологии Big Data опираются на вычислительные кластеры из множества вычислителей, снабженных локальной подсистемой хранения. Доступ к данным и их обработка осуществляются специальным программным обеспечением. Наиболее известным и интенсивно развивающимся проектом в области Big Data является Apache Hadoop[2].

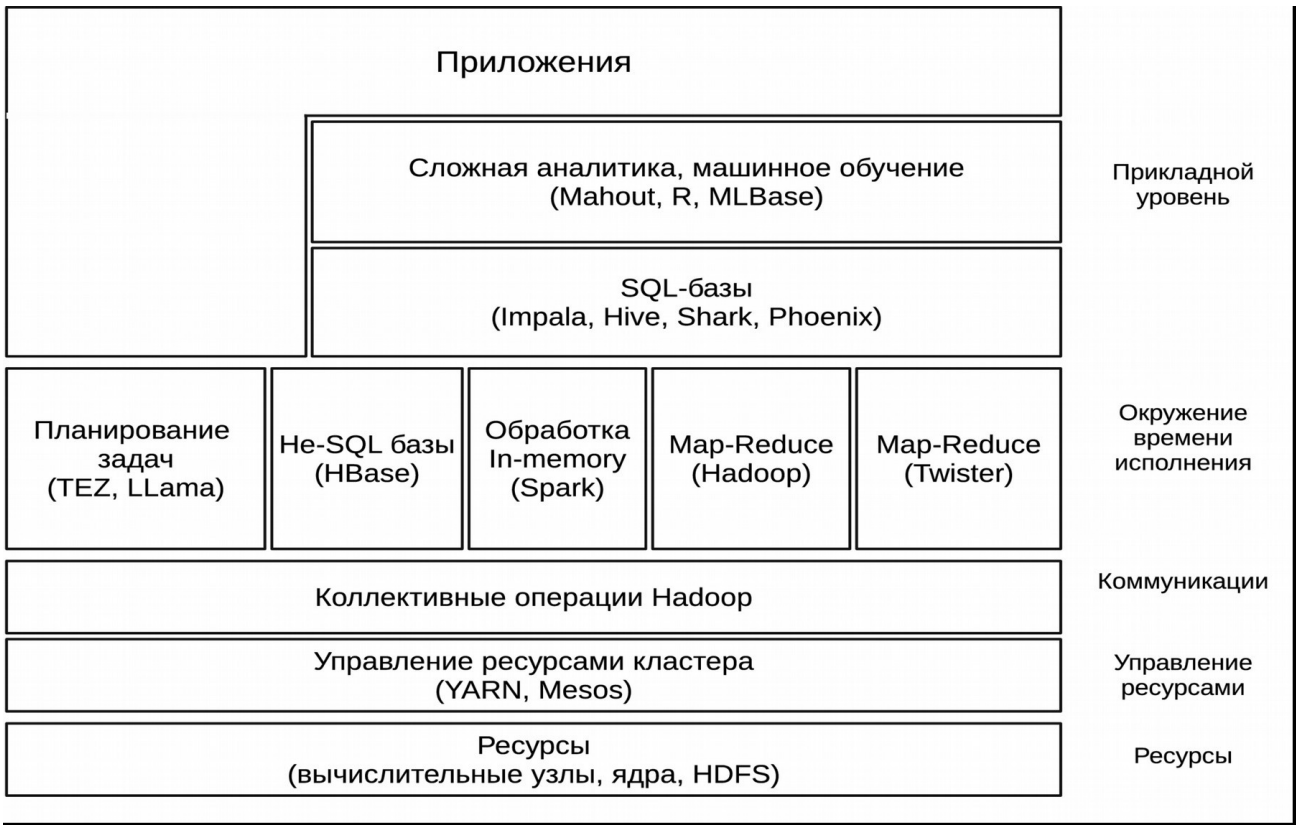

Рис. 1: Стек технологий Big Data Apache (ABDS)

Изначально, в проекте развивались два взаимосвязанных направления: распределенная файловая система HDFS (Hadoop Distributed File System) и система вычислений по методу Map-Reduce. К настоящему времени на базе Apache Hadoop был создан стек продуктов Big Data, получивший название Apache Big Data Stack, или сокращенно ABDS (рисунок 1). В этом стеке насчитывается более 110 проектов различного назначения[3].

В зависимости от прикладной задачи разработчик приложения может воспользоваться продуктами прикладного уровня, или непосредственно использовать интерфейс окружения времени исполнения. Уровни коммуникаций и управления ресурсами задействуются автоматически.

Курс «Высокопроизводительные распределенные системы» ориентирован на изучение открытого программного обеспечения, входящего в стек технологий ABDS.

Доступность программного обеспечения позволяет слушателям курса самостоятельно устанавливать необходимые инструменты и утилиты и экспериментировать с технологией.

2 Практические занятия

Практические занятия проходят в компьютерных классах с предустановленным программным обеспечением:

1. Oracle Virtual Box
2. Oracle Java 1.7
3. Intellij IDEA 15 Community Edition
4. Apache Hadoop
5. Apache Spark
6 Apache R
7. Apache Hive
8. Apache Hbase
9. Apache Tomcat

2.1 Занятие №1 Параллельное программирование в Java

Цель: ознакомление с современными компонентами параллельной многопоточной (multithread) обработки данных на платформе Java 1.7

Вопросы подлежащие рассмотрению:

1. Оргинизация проекта на языке Java в среде IntelliJ IDEA
2. Запуск и исследование примеров использования компонентов параллельной обработки: Thread, Callable, Future, FutureTask, пулы потоков, механизмы fork/join
3. Решение задач на применение инструментов параллельной обработки данных Java

Рекомендуемая литература:

1. Java - новое поколение разработки / Эванс В. - СПб. : ПИТЕР, 2014. - 560с.

2.2 Занятие №2 Технология Hadoop и Map/Reduce

Цель: ознакомление с технологиями Hadoop и Map/Reduce, овладение программным и системным инструментарием обработки данных в парадигме Big Data

Вопросы подлежащие рассмотрению:

1. Конфигурирование и системные утилиты Hadoop, взаимодействие с файловой системой HDFS

2. Конфигурирование и системные утилиты Map/Reduce, запуск примеров программ обработки данных

3. Разработка собственной программы на языке Java для Map/Reduce Hadoop

Рекомендуемая литература:

1. T. Whyte, Hadoop: The Definitive Guide, 4th Edition 2015

2.3 Занятие №3 Технология Apache Spark + Java

Цель: ознакомление с технологией Apache Spark, овладение программным обеспечение обработки данных в системе Spark на языке Java

Вопросы подлежащие рассмотрению:

1. Конфигурирование и системные утилиты Apache Spark, взаимодействие с классическими и распределенными файловыми системами

2. Запуск примеров программ в системе Spark на языке Java. Измерение производительности

3. Разработка собственной программы на языке Java для Apache Spark

Рекомендуемая литература:

1. Изучаем Spark / Захария М., Венделл П., Конвински Э., Карау Х. - Москва: ДМК-Пресс, 2015. - 400с.

2.4 Занятие №4 Технология Apache Spark + R

Цель: ознакомление с языком анализа данных R, овладение программным обеспечением анализа данных в системе Spark на языке R

Вопросы подлежащие рассмотрению:

1. Интерактивная среда на языке R в системе Apache Spark

2. Запуск примеров программ в системе Spark на языке R. Измерение производительности

3. Разработка собственной программы на языке R для Apache Spark

Рекомендуемая литература:

1. Шипунов А.Б., Балдин Е.М. [Электронный ресурс] – режим доступа: http://www.inp.nsk.su/~baldin/DataAnalysis/index.html — свободный

2.5 Занятие №5 Технологии Apache Hive, Hbase

Цель: ознакомление с реляционными и не-реляционными технологиями хранения данных Apache Hive и HBase, овладение программным обеспечением создания, поиска и изменения данных в системах Hive и HBase

Вопросы подлежащие рассмотрению:

1. Запуск примеров программ для Apache HBase. Измерение производительности

2. Запуск примеров программ для Apache Hive. Измерение производительности

3. Разработка собственных программ на языке Java для Apache Hive и HBase

Рекомендуемая литература:

1. Изучаем Spark / Захария М., Венделл П., Конвински Э., Карау Х. - Москва: ДМК-Пресс, 2015. - 400с.

3 Самостоятельная работа

3.1 Темы на самостоятельное изучение

Таблица 1 - Темы выносимые на самостоятельное изучение

Тема	Трудоемкость, ч.	Рекомендуемая литература
Технологии высокопроизводительных вычислений MPI, OpenMP	2	[8]
Задачи на Map/Reduce	4	[5, 7]
Распределенные файловые системы	4	[5, 7]
Области применения ABDS	2	[9]

3.2 Темы рефератов

1. Система Hive
2. Система Impala
3. Система Shark
4. Система Phoenix
5. Сравнение технологий MapReduce: Hadoop и Twister
6. Среда R
7. Система Mahout
8. Система MLBase
9. Файловая система MapR
10. Системы планирования задач
11. Система YARN
12. Система Mesos

Литература

1. Demchenko, Y. Defining architecture components of the Big Data Ecosystem / Y. Demchenko, C. De Laat, P. Membrey // Collaboration Technologies and Systems (CTS), 2014 International Conference on. — 2014. — May. — P. 104–112.

2. Проект Apache Hadoop [Электронный ресурс]. — Режим доступа: https://hadoop.apache.org/, свободный.

3. A Tale of Two Data-Intensive Paradigms: Applications, Abstractions, and Architectures / S. Jha, J. Qiu, A. Luckow et al. // 2014 IEEE International Congress on Big Data. — 2014. — Jun.

4. Java - новое поколение разработки / Эванс В. - СПб. : ПИТЕР, 2014. - 560с.

5. Tome Whyte Hadoop: The Definitive Guide, 4th Edition 2015

6. Шипунов А.Б., Балдин Е.М. [Электронный ресурс] – режим доступа: http://www.inp.nsk.su/~baldin/DataAnalysis/index.html — свободный

7. Изучаем Spark / Захария М., Венделл П., Конвински Э., Карау Х. - Москва: ДМК-Пресс, 2015. - 400с.

8. Современные языки и технологии параллельного программирования [Текст] : учебник для вузов / В. П. Гергель; авт. предисл. В. А. Садовничий; Библиотека Нижегородского государственного университета имени Н. И. Лобачевского (Нижний Новгород). - М. : Издательство Московского университета, 2012. - 408 с. : ил. - (Суперкомпьютерное образование). - Библиогр.: с. 394-402. - ISBN 978-5-211-06380-8

9. S. Ryza, U. Laserson, S. Owen, J. Wills, Advanced Analytics with Spark, 2015