УДК 004.633.2

АНАЛИЗ ПРОИЗВОДИТЕЛЬНОСТИ ТЕХНОЛОГИИ HADOOP

А.А. Беляк
студент БГУИР

С.Н. Нестеренков
Доцент кафедры программного обеспечения информационных технологий, Кандидат технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, факультет. Компьютерного проектирования, кафедра проектирования информационно-компьютерных систем, Республика Беларусь.
E-mail: alexbeljak99@gmail.com, s.nesterenkov@bsuir.by.

А. А. Беляк
Студент 4 курса специальности «Программируемые мобильные системы» БГУИР.
С. Н. Нестеренков
Окончил БГУИР в 2007 году по специальности «Программное обеспечение информационных технологий», окончил магистратуру БГУИР в 2008 по специальности «Системный анализ, управление и обработка информации», окончил аспирантуру БГУИР в 2013 по специальности «Системный анализ, управление и обработка информации», окончил магистратуру БГУИР в 2013 по специальности «Экономика и управление народным хозяйством», в 2017 защитил диссертацию на соискание ученой степени кандидата технических наук по специальности «Системный анализ, управление и обработка информации».

Аннотация. Описаны основные технологии Big Data, рассмотрен Hadoop, проанализирована производительность Hadoop при сортировке файлов различного размера. В результате получено, что увеличение объёма данных подразумевает увеличение времени выполнения, получены оптимальные настройки узлов кластеров в зависимости от объёма входных данных. Приведены диаграммы сравнения производительности Hadoop в зависимости от размера файлов.

Ключевые слова: Hadoop, оценка производительности Hadoop, технологии Big Data, TeraSort Benchmark, TestDFSIO Benchmark.

Введение.

В ходе своей работы в глобальной сети интернет каждый пользователь производит большое число данных, которые требуется хранить и правильно обрабатывать. Для работы с этими данными необходимо все больше ресурсов и увеличение сложности программных решений. Для обозначения больших объемов информации, а также технологий поиска, обработки и применения неструктурированных данных в больших объемах существует термин «Big Data». Анализу производительности основных технологий Big Data и посвящена данная работа. Весь анализ было решено производить с использованием Hadoop.

Описание Hadoop.

Hadoop является одной из основополагающих технологий Big Data. Hadoop – проект фонда Apache Software Foundation, свободно распространяемый набор утилит, библиотек и фреймворк для разработки и выполнения распределённых программ, работающих на кластерах из сотен и

тысяч узлов [1, 2]. Применяется для реализации поисковых и контекстных механизмов многих веб-сайтов с высокой нагрузкой. Среди наиболее известных можно выделить Yahoo! и Facebook [3]. Разработан на языке программирования Java в рамках парадигмы MapReduce. Согласно данной парадигме, приложение делится на большое количество равнозначных элементарных заданий, выполняемых на узлах кластера и сводимых в конечный результат [4, 5].

Среди основных модулей Hadoop можно выделить [6].

1) HDFS – это распределённая файловая, предназначенная для работы на стандартном оборудовании. В сравнении с другими распределёнными файловыми системами HDFS отличается высокой отказоустойчивостью и предназначена для развертывания на недорогом оборудовании. HDFS обеспечивает высокопроизводительный доступ к данным приложения и подходит для приложений с большими наборами данных.

2) YARN. Фундаментальная идея – разделить функции управления ресурсами и планирования или мониторинга заданий на отдельные потоки-демоны. Идея состоит в том, чтобы иметь глобальный ResourceManager (RM) и ApplicationMaster (AM) для каждого приложения. Приложение – это либо отдельное задание, либо группа заданий [7, 8].

3) MapReduce – это программная среда для простого написания приложений, которые обрабатывают огромные объемы данных (многотерабайтные наборы данных) параллельно на больших кластерах стандартного оборудования надежным и отказоустойчивым способом.

Анализ производительности Hadoop.

Использованный для анализа экспериментальный кластер состоит из 6 узлов. Один из них предназначен для работы в качестве главного узла. Остальные 5 узлов предназначены для работы в качестве подчинённых узлов. Информация об оборудовании каждого узла следующая: 8 ядер, 15 гигабайт оперативной памяти, 80 гигабайт SSD,

Для анализа производительности сортировки использовался TeraSort Benchmark. Он используется для тестирования как MapReduce, так и HDFS путем максимально быстрой сортировки некоторого объема данных, чтобы измерить возможности распределения и сопоставления файлов в кластере. Этот тест состоит из трёх компонентов: TeraGen (генерирует случайные данные), TeraSort (сортирует данные, используя MapReduce), TeraValidate (используется для проверки выходных данных).

Для анализа производительности чтения и записи использовался TestDFSIO Benchmark. Он используется для тестирования производительности чтения и записи HDFS. Данная утилита используется для таких задач, как стресс-тестирование HDFS, чтобы обнаружить места с плохой производительностью, также она позволяет дать первое впечатление о том, как быстро работает кластер с точки зрения ввода или вывода.

TestDFSIO разработан таким образом, что он будет использовать 1 задачу карты для каждого файла, то есть это отображение один к одному из файлов в задачи карты. Разделения определены так, что каждая карта получает только одно имя файла, которое она создает (-write) или читает (-read).

Для генерации случайных данных использовалась команда hadoop jar hadoop-test*test*.jar TestDFSIO -write|-read -nrFiles <no. of output files> -fileSize <size of one file>. Для сортировки сгенерированных данных использовалась команда hadoop jar $HADOOP_HOME/Hadoop-*examples*.jar terasort <input dir> <output dir>.

Оценка производительности Hadoop на созданном кластере проводилась с помощью 5 файлов размером 10, 20, 30, 40 и 50 гигабайт соответственно Сравнительный анализ генерации случайных данных, их сортировке и проверке выходных данных приведён на рисунке 3 Оценка производительности Hadoop приведена на рисунке 4 Оценка производительности чтения/записи модуля HDFS приведена на рисунке 5.

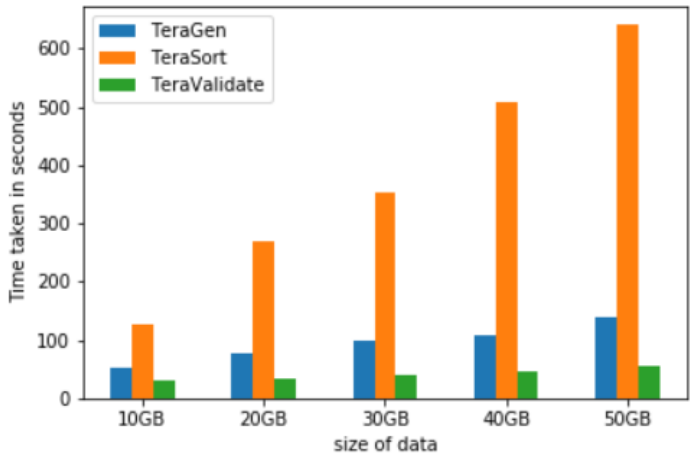

Рисунок 1. Сравнительный анализ генерации, сортировки и валидации выходных данных при использовании TeraSort

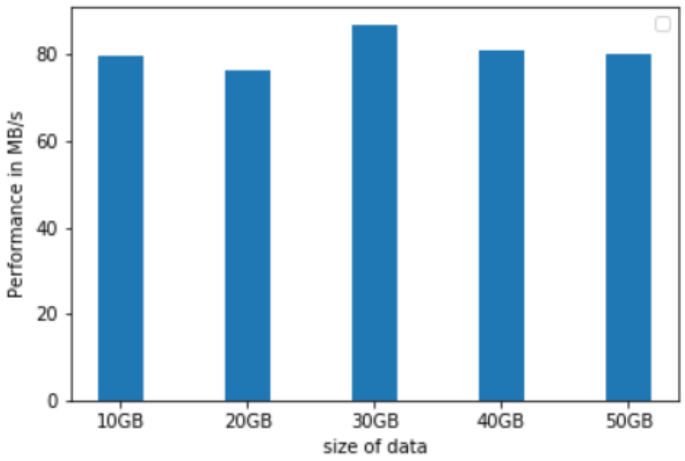

Рисунок 2. Результаты сравнительного анализа производительности Hadoop путём сортировки файлов различного размера

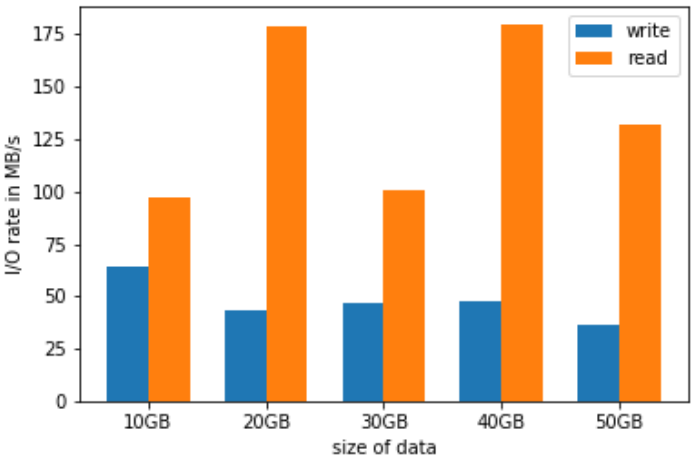

Рисунок 3. Результаты анализа производительности чтения/записи модуля HDFS на файлах различного размера

Заключение.

Исходя из приведённого выше сравнительного анализа можно сделать следующие выводы.

1. Увеличение объёма данных для сортировки подразумевает увеличение времени выполнения.

2. Для средних размеров данных, таких как 30 или 40 гигабайт наилучшим по производительности является кластер из 1 ведущего и 5 ведомых узлов.

3. Для больших размеров данных кластер из 1 ведущего и 5 ведомых узлов не является наилучшим по производительности.

Список литературы

[1] Martin K. Designing Data-Intensive Applications. – O'Reilly Media, 2017. - 616 p.

[2] James W. Big Data: Principles and best practices of scalable realtime data systems. – Manning Publications, 2015. - 328 p.

[3] Tom W. Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale 4th Edition. – O'Reilly Media, 2015. - 756 p.

[4] Нестеренков, С.Н. Применение больших данных в электронном образовании / С.Н. Нестеренков, М.И. Макаров, Н.В. Ющенко, А.Д. Радкевич // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня : сб. материалов V Междунар. науч.-практ. конф. (Республика Беларусь, Минск, 13-14 марта 2019 года). В 2 ч. Ч. 2 / редкол. : В. А. Богуш [и др.]. - Минск : БГУИР, 2019. - С. 242-245.

[5] Калоша, А.Л. Система анализа качества текстовых коллекций / А.Л. Калоша, М.А. Медунецкий, М.П. Хоронеко, А.А. Александров, А.И. Гридасов, С.Н. Нестеренков // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня : сб. материалов VI Междунар. науч.-практ. конф. (Республика Беларусь, Минск, 20-21 мая 2020 года): в 3 ч. Ч. 2 / редкол. : В. А. Богуш [и др.]. - Минск : Бестпринт, 2020. - С. 369-375.

[6] Bernard M. Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance. – Wiley, 2015. - 256 p.

[7] Кукареко, А.В. Способы машинного обучения для выявления ошибок выполнения упражнений на smart-тренажере / А.В. Кукареко, С.Н. Нестеренков // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня : сб. материалов VI Междунар. науч.-практ. конф. (Республика Беларусь, Минск, 20-21 мая 2020 года): в 3 ч. Ч. 2 / редкол. : В. А. Богуш [и др.]. - Минск : Бестпринт, 2020. - С. 214-224.

[8] Зязюлькин, С. П. Использование DQN для обучения агентов игр (Atari 2600)/ С. П. Зязюлькин, С. Н. Нестеренков // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня : сб. материалов VI Междунар. науч.-практ. конф. (Республика Беларусь, Минск, 20-21 мая 2020 года): в 3 ч. Ч. 2 / редкол. : В. А. Богуш [и др.]. – Минск : Бестпринт, 2020. – С. 274-280.

PERFORMANCE ANALYSIS OF HADOOP TECHNOLOGY

A.A. BELIAK
Student of Belarusian State University of Informatics and Radioelectronics

S.N. NESTERENKOV,
PhD Associate professor of department of the software of information technologies

Belarusian State University of Informatics and Radioelectronics
E-mail: alexbeljak99@gmail.com, s.nesterenkov@bsuir.by

Abstract. The main Big Data technologies are described, Hadoop is considered, the performance of Hadoop is analyzed when sorting files of various sizes. As a result, the obtained data obtained the optimal settings of the cluster nodes, depending on the amount of input data. There are diagrams comparing Hadoop performance versus file size.

Keywords: Hadoop, Hadoop performance evaluation, Big Data technologies, TeraSort Benchmark, TestDFSIO Benchmark.