УДК 004

ИСПОЛЬЗОВАНИЕ МЕТОДА MAPREDUCE В BIG DATA

А. А. Некратюк, О. А. Сафарьян
Донской государственный технический университет (г. Ростов-на-Дону, Российская Федерация)

Задача работы — создать программное средство, реализующее модель обработки больших данных с помощью алгоритма MapReduce. Рассмотрена теория обработки и использования больших данных, показаны примеры листингов функций, необходимых для дальнейшей работы.
Ключевые слова: анализ данных, большие данные, MapReduce, база данных, алгоритм, value, velocity, variety.

USE OF MAPREDUCE METHOD IN BIG DATA

A. A. Nekratyuk, O. A. Safaryan
Don State Technical University (Rostov-on-Don, Russian Federation)

In this paper, the task is to create a software tool that implements a model for processing big data using the MapReduce algorithm. The paper provides general information about the theory of processing and the use of big data, shows examples of function listings necessary for further work.
Keywords: data analysis, big data, MapReduce, database, algorithm, value, velocity, variety.

Введение. Сегодня почти вся информация, с которой работает человек, хранится в электронном виде и с каждым днем объем данных увеличивается. С помощью различных методов анализа можно выявлять закономерности, скрытые в этих данных, например, извлекать полезную коммерческую информацию для оптимизации бизнес-процессов или даже предвидеть катастрофу.

Большие данные (Big Data) — это собирательное обозначение структурированных и неструктурированных данных огромных объёмов, отличающихся большим разнообразием, которые можно эффективно обрабатывать программными инструментами. Big Data является социально-экономическим феноменом, связанным с появлением технических возможностей обрабатывать и анализировать огромные массивы данных в различных сферах человеческой деятельности.

Для обработки и анализа Big Data используются разные инструменты и подходы, одним из самых эффективных является метод MapReduce. Это модель проведения распределённых вычислений, разработанная компанией Google. Данная схема используется в технологии обработки больших наборов данных в компьютерных кластерах [1, 2].

Постановка задачи. Задачей исследования является попытка создания модели программного средства для сортировки больших объемов данных на примере системы управления реляционными базами данных MS SQL Server.

Теоретические сведения. Big data представляет собой любые данные такого объема, что их нельзя обработать на одном компьютере. Термин Big Data определяется как данные объема, измеряемого в терабайтах. В качестве определяющих характеристик больших данных можно выделить «три V»: Value — объём данных; Velocity (скорость) — насколько быстро возрастает объем данных и насколько быстро их можно обрабатывать, Variety (разнообразие) — с данными каких типов необходимо работать [2] (рис 1).

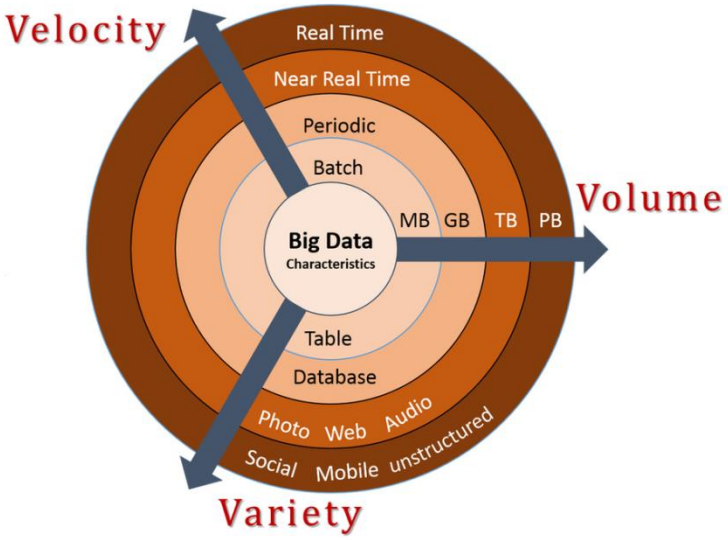


Рис. 1. Схема характеристик больших данных

Развитие технологий и популярность социальных сетей способствуют экспрессному росту объемов данных, генерируемых как людьми, так и машинами. Они распространяются в различных местах и форматах в огромных объемах и с огромной скоростью. Признак скорости — это быстрота генерации и прироста данных. Получение необходимых данных в кратчайшие сроки является важным конкурентным преимуществом для принятия решений, потому что разные ситуации имеют различные требования к временным промежуткам, в течение которых это решение будет актуально [2]. Признак разнообразия относится к различным форматам хранения данных.

Человечество генерирует огромные объемы структурированных и неструктурированных данных. До начала активного развития технологий Big Data не было достаточно мощных и эффективных инструментов, способных обрабатывать такие объемы неструктурированных данных, с которыми приходится работать сегодня [2].

Потребление и обработка огромного количества данных является необходимостью для организаций в современном мире, стремящихся сохранить конкурентоспособность. К данным Big data относятся не только структурированные электронные таблицы, но и неструктурированная информация: изображения, аудио и видеофайлы, данные, собранные с физических датчиков приборов и многое другое [3]. В табл. 1 представлена сравнительная характеристика традиционной базы и базы Big Data.

Таблица 1

Сравнительные характеристики данных

Характеристика	**Традиционная база данных**	**База больших данных**
Объем информации	От гигабайт (10^9 байт) до терабайт (10^{12} байт)	От петабайт (10^{15} байт) до эксабайт (10^{18} байт)
Способ хранения	Централизованный	Децентрализованный
Структурированность данных	Структурирована	Полуструктурирована и неструктурирована
Модель хранения и обработки данных	Вертикальная модель	Горизонтальная модель
Взаимосвязь данных	Сильная	Слабая

Существуют отрасли, данные в которых собираются и накапливаются более интенсивно, чем в других. Для ситуаций, в которых есть необходимость хранения данных годами, накопленная информация классифицируется как экстремально большие данные [3]. В табл. 2 приведена сравнительная характеристика частоты обработки информации различных типов для разных сфер деятельности. Здесь обозначение степени использования: L — низкая, M — средняя, H — высокая.

Таблица 2

Степень использования данных в разных сферах деятельности

	Видео	Изображения	Текст/числа
Банковский сектор	M	M	H
Страхование	L	L	H
Ценные бумаги и инвестиции	L	L	H
Производство	M	M	H
Розничная торговля	M	L	H
Оптовая торговля	L	L	H
Профессиональные услуги	M	M	H
Развлекательные услуги	M	L	M
Здравоохранение	L	H	H
Транспортные услуги	M	M	H
СМИ	H	M	H
Коммунальные услуги	M	M	H

Так же растет количество экстремально больших данных в коммерческих и государственных секторах, объем данных такого рода находится в хранилищах и зачастую составляет сотни петабайт [3]. Для обработки таких объемов существует множество алгоритмов, одним из наиболее эффективных является MapReduce. Работа этого алгоритма состоит из пошаговых функций:

1. Map — предварительная обработка и сортировка входных данных. При этом главный узел кластера получает значения этих данных, делит их на части и распределяет по рабочим узлам. Каждый рабочий узел сохраняет результат в формате «ключ — значение» во временном хранилище.

2. Shuffle — рабочие узлы перераспределяют данные на основе ключей, созданных функцией Map, таким образом, что все данные одного ключа попадают в один рабочий узел.

3. Reduce — одновременная обработка данных каждым рабочим узлом всех групп по порядку следования ключей и сбор результатов. Главный узел получает промежуточные результаты от рабочих узлов и передаёт их на свободные узлы для выполнения следующего шага. Итоговый результат является решением исходной задачи [4, 5]. На рис. 2 приведена схема работы алгоритма MapReduce.

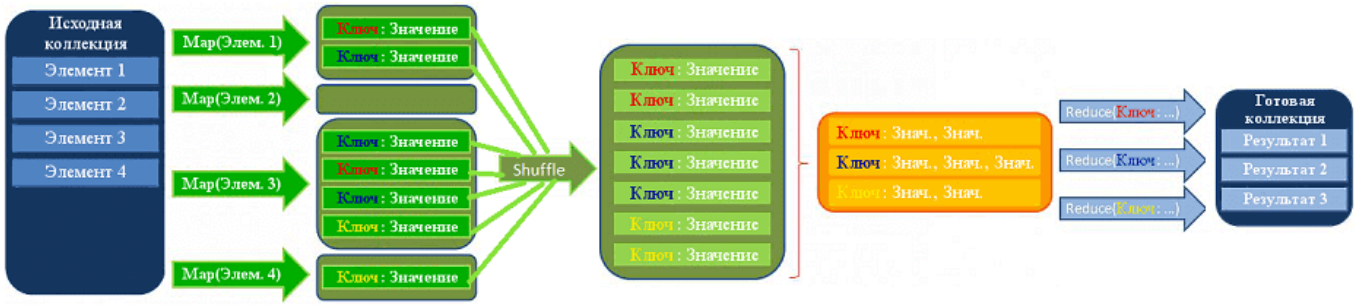


Рис. 2. Схема работы алгоритма MapReduce

Описание программного средства. Разрабатываемое программное средство выполняет функции:

- подключение к базе данных;
- обработка и сортировка информации с использованием метода MapReduce;
- вывод результата сортировки данных.

На текущий момент реализованы все элементы программного средства, необходимые для работы алгоритма MapReduce. Основная проблема в решении данной задачи — очень большие объемы данных. Так как метод MapReduce является алгоритмом поиска по всей информации в базе без индексирования входных данных, то для обработки действительно больших объемов требуются вычислительные мощности, сравнимые с суммарной мощностью десятков среднестатистических компьютеров. На рис. 3, приведены фрагменты кода реализации алгоритма MapReduce.

```
USE [Travmpunkt]
GO
/****** Object:  View [dbo].[Crime]    Script Date: 12/11/2019 22:38:39 ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
create view [dbo].[Crime] as SELECT * from
journal where mark = 4
GO
USE [Travmpunkt]
GO
/****** Object:  View [dbo].[DTP_M_18_40]    Script Date: 12/11/2019 22:38:45 ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
create view [dbo].[DTP_M_18_40] as SELECT * from
DTP_M where age between 18 and 40
GO
USE [Travmpunkt]
GO
/****** Object:  View [dbo].[Svodka]    Script Date: 12/11/2019 22:38:51 ******/
SET ANSI_NULLS ON
GO
```

Рис. 3. Создание таблиц-представлений

Алгоритм реализуется за счет создания таблиц-представлений, которые сортируют данные соответственно сортировке на прошедшем этапе. Схема алгоритма представлена на рис. 4.

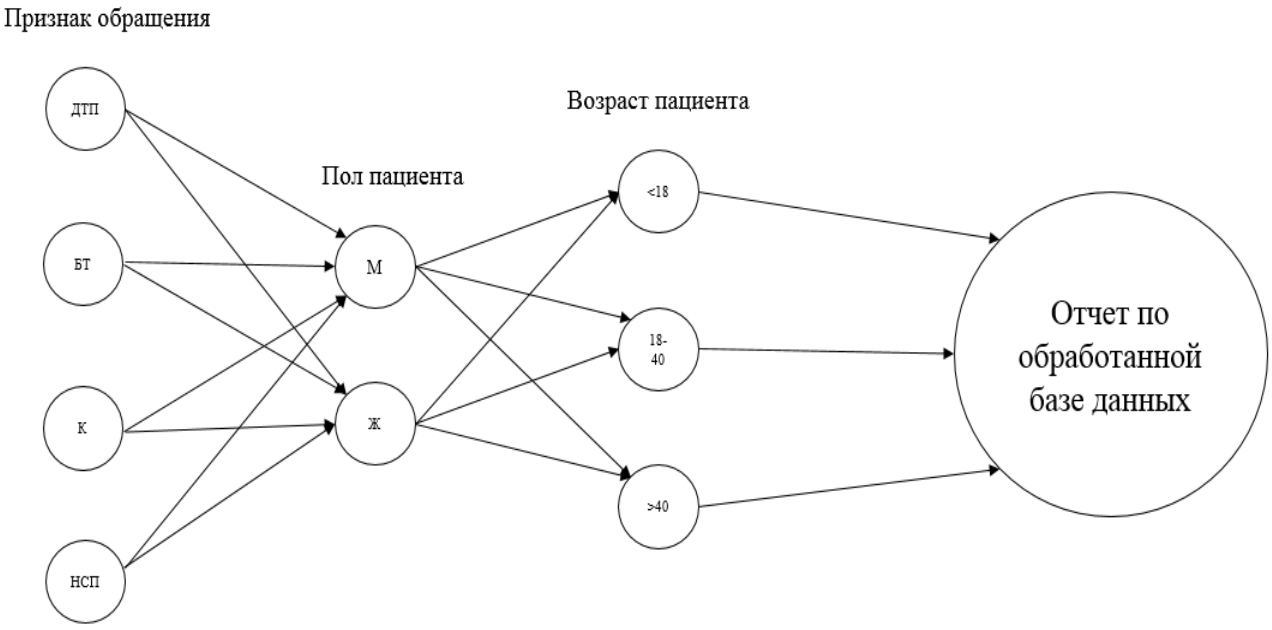
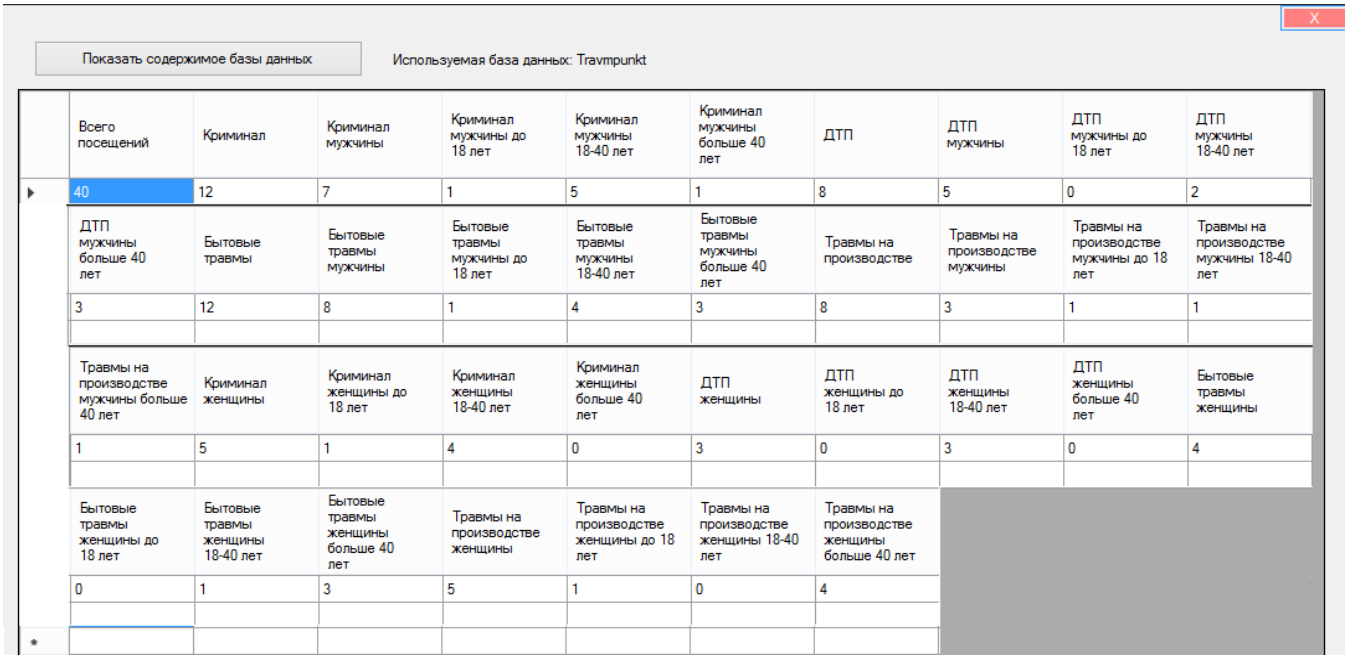


Рис. 4. Схема работы алгоритма MapReduce

В результате выполнения алгоритма на тестовой базе были получены результаты, представленные на рис. 7.



Всего посещений	Криминал	Криминал мужчины	Криминал мужчины до 18 лет	Криминал мужчины 18-40 лет	Криминал мужчины больше 40 лет	ДТП	ДТП мужчины	ДТП мужчины до 18 лет	ДТП мужчины 18-40 лет
40	12	7	1	5	1	8	5	0	2
ДТП мужчины больше 40 лет	Бытовые травмы	Бытовые травмы мужчины	Бытовые травмы мужчины до 18 лет	Бытовые травмы мужчины 18-40 лет	Бытовые травмы мужчины больше 40 лет	Травмы на производстве	Травмы на производстве мужчины	Травмы на производстве мужчины до 18 лет	Травмы на производстве мужчины 18-40 лет
3	12	8	1	4	3	8	3	1	1
Травмы на производстве мужчины больше 40 лет	Криминал женщины	Криминал женщины до 18 лет	Криминал женщины 18-40 лет	Криминал женщины больше 40 лет	ДТП женщины	ДТП женщины до 18 лет	ДТП женщины 18-40 лет	ДТП женщины больше 40 лет	Бытовые травмы женщины
1	5	1	4	0	3	0	3	0	4
Бытовые травмы женщины до 18 лет	Бытовые травмы женщины 18-40 лет	Бытовые травмы женщины больше 40 лет	Травмы на производстве женщины	Травмы на производстве женщины до 18 лет	Травмы на производстве женщины 18-40 лет	Травмы на производстве женщины больше 40 лет			
0	1	3	5	1	0	4			

Рис. 5. Результат работы алгоритма

Таким образом, алгоритм работает корректно и позволяет последовательно анализировать данные, находящиеся в обрабатываемой базе.

Заключение. В настоящее время проводится работа над программным средством по динамической сортировке больших данных с помощью метода MapReduce. Успешно реализована ключевая часть алгоритма в виде промежуточных таблиц-представлений. В дальнейшем будут реализованы функции для подключения к базам данных с динамической структурой. Для работы алгоритма база должна быть строго структурирована.

Библиографический список

1. Nathan Martz, James Warren. Big Data: Principles and Best Practices of Scalable Realtime Data Systems. 2017 – 8с.
2. Viktor Mayer-Schönberger, Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work, and Think. 2014 – 38 с.
3. Rick Smolan, Jennifer Erwitt. The Human Face of Big Data. 2012 – 12 с.
4. Thomas H. Davenport. Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. 2014 – 48 с.
5. Miner Donald, Shook Adam. MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems. 2012 – 18 с.

Об авторах:

Некратюк Антон Анатольевич, студент кафедры «Кибербезопасность информационных систем» Донского государственного технического университета (344000, РФ, г. Ростов-на-Дону, пл. Гагарина, 1), antonnekratyuk@yandex.ru

Сафарьян Ольга Александровна, доцент кафедры «Кибербезопасность информационных систем» Донского государственного технического университета (344000, РФ, г. Ростов-на-Дону, пл. Гагарина, 1), кандидат технических наук, доцент, safari_2006@mail.ru

Authors:

Nekratyuk, Anton A., student of the Department of Cybersecurity of Information Systems, Don State Technical University (1, Gagarin sq., Rostov-on-Don, 344000, RF), antonnekratyuk@yandex.ru

Safaryan, Olga A., associate professor of the Department of Cybersecurity of Information Systems, Don State Technical University (1, Gagarin sq., Rostov-on-Don, 344000, RF), Cand.Sci., associate professor, safari_2006@mail.ru