**БАЛТИЙСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ "ВОЕНМЕХ"**

**им. Д. Ф. УСТИНОВА**

***Кафедра И9***

**ОТЧЕТ ПО НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ**

магистранта Николаева Антона Андреевича

группы И9М31

**ПРЕПОДАВАТЕЛЬ**

Верхолат А.М. / \_\_\_\_\_\_\_\_\_\_\_\_\_\_ /

Фамилия И.О. Подпись

“\_\_\_" \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ***2018 г.***

САНКТ-ПЕТЕРБУРГ

2018 г.

Содержание

[1 Аннотация научно-исследовательской работы в семестре 3](#_Toc5281860)

[2 Основные итоги выполнения курсовой работы 3](#_Toc5281861)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 5](#_Toc5281862)

# 1 Аннотация научно-исследовательской работы в семестре

В соответствии с планом научно-исследовательской работы на семестр была написана курсовая работа, обозревающая основные технологии, которые позволяют эффективно обрабатывать неструктурированные данные и выбрана наиболее удачная для разработки тестового проекта в рамках магистерской диссертации. Исследованы различные алгоритмы обработки, полученных знаний после первичной обработки, а также подходы к хранению таких данных. Изучены NOSQL и реляционные подходы к построению базы данных.

# 2 Основные итоги выполнения курсовой работы

Цель курсовой работы - исследование основных инструментов хранения и обработки неструктурированной информации и выбор наиболее эффективного для дальнейшего использования в рамках выполнения выпускной квалификационной работы.

Для достижения поставленной цели решаются следующие задачи:

* Исследованы различные подходы к извлечению информации;
* Изучено понятие Big Data и его взаимосвязь с неструктурированными данными;
* Изучены способы распределенной обработки различных типов данных;
* Изучены такие технологии как Apache Hadoop, Apache Spark и Apache Tez;
* Выбраны наиболее эффективные технологии для тестого проекта диссертации.

В ходе выполнения курсовой работы был проведен анализ самых распространенных технологий в области больших данных для обработки слабоструктурированных и неструктурированных данных. Результаты, полученные в ходе анализа, позволяют легко сделать выбор в пользу комбинации из слабоструктурированного формата данных и одной, наиболее оптимальной в рассматриваемом случае, технологии - Apache Spark, как части Hadoop-кластера. Кроме технических достоинств немаловажны и социально-коммерческие. Поэтому стоит обратить внимание и на то, что проект Apache Spark привлекает к себе огромное внимание, про него написано большое количество маленьких практических статей, он стал частью Hadoop 2.0. Плюс он быстро оброс дополнительными фреймворками, такими, как Spark Streaming, SparkML, Spark SQL, GraphX, а кроме этих «официальных» фреймворков появилось множество проектов — различные коннекторы, алгоритмы, библиотеки и так далее. Что говорит о его заслуженной оценённости в мире IT и соответствии ожиданиям в контексте рассматриваемого проекта. Так же было обнаружено что для полноты анализа необходима вторичная обработка данных, полученных после обработки на Hadoop-кластере, с помощью специализированных СУБД как классических реляционных, так и NOSQL решений. В зависимости от задачи и самих данных выбор такой СУБД будет меняться.

Таким образом, задачи, поставленные в курсовой работе, были решены.

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Tom White. Hadoop: The Definitive Guide: Storage and Analysis at internet Scale. –: "O'Reilly Media Inc. ", 2015. – 756 с.
2. Чак Лэм. Hadoop в действии. –: "Litres", 2017.
3. [Srinath Perera.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Srinath+Perera%22&source=gbs_metadata_r&cad=7) Hadoop Mapreduce Cookbook. –: "Packt Publishing Ltd", 2013. – 300 c.
4. [VK Jain.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22VK+Jain%22&source=gbs_metadata_r&cad=5) Big Data and Hadoop. –: "Khanna Publishing", 2017. – 600 c.
5. [Jason Venner,](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Jason+Venner%22&source=gbs_metadata_r&cad=6) [Sameer Wadkar,](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Sameer+Wadkar%22&source=gbs_metadata_r&cad=6) [Madhu Siddalingaiah.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Madhu+Siddalingaiah%22&source=gbs_metadata_r&cad=6) Pro Apache Hadoop. –: "Apress", 2014. – 444 c.
6. [Arun Murthy,](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Arun+Murthy%22&source=gbs_metadata_r&cad=7) [Vinod Vavilapalli,](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Vinod+Vavilapalli%22&source=gbs_metadata_r&cad=7) [Douglas Eadline,](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Douglas+Eadline%22&source=gbs_metadata_r&cad=7) [Joseph Niemiec,](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Joseph+Niemiec%22&source=gbs_metadata_r&cad=7) [Jeff](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Jeff+Markham%22&source=gbs_metadata_r&cad=7) [Markham.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Jeff+Markham%22&source=gbs_metadata_r&cad=7) Apache Hadoop YARN: Moving beyond MapReduce and Batch Processing with Apache Hadoop 2. –: "Addison–Wesley Professional", 2014. – 400 c.
7. [Dirk deRoos.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Dirk+deRoos%22&source=gbs_metadata_r&cad=7) Hadoop For Dummies. –: "John Wiley & Sons", 2014. – 408 c.
8. [Vignesh Prajapati.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Vignesh+Prajapati%22&source=gbs_metadata_r&cad=7) Big Data Analytics with R and Hadoop. –: "Packt Publishing Ltd", 2013. – 238 c.
9. [Boris Lublinsky,](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Boris+Lublinsky%22&source=gbs_metadata_r&cad=7) [Kevin T. Smith,](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Kevin+T.+Smith%22&source=gbs_metadata_r&cad=7) [Alexey Yakubovich.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Alexey+Yakubovich%22&source=gbs_metadata_r&cad=7) Professional Hadoop Solutions. –: "John Wiley & Sons", 2013. – 504 c.
10. [Kevin Sitto,](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Kevin+Sitto%22&source=gbs_metadata_r&cad=7) [Marshall Presser.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Marshall+Presser%22&source=gbs_metadata_r&cad=7) Field Guide to Hadoop: An Introduction to Hadoop, Its Ecosystem, and Aligned Technologies. –: "O'Reilly Media, Inc.", 2015. – 132 c.
11. [Garry Turkington.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Garry+Turkington%22&source=gbs_metadata_r&cad=6) Hadoop Beginner's Guide. –: "Packt Publishing Ltd", 2013.-398 c.
12. [Danil Zburivsky.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Danil+Zburivsky%22&source=gbs_metadata_r&cad=7) Hadoop Cluster Deployment. –: "Packt Publishing Ltd", 2013. – 126 c.

1. [Kevin Roebuck.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Kevin+Roebuck%22&source=gbs_metadata_r&cad=2) MapReduce: High–impact Strategies – What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors. –: "Lightning Source", 2011. – 170 c.
2. [Donald Miner,](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Donald+Miner%22&source=gbs_metadata_r&cad=7) [Adam Shook.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Adam+Shook%22&source=gbs_metadata_r&cad=7) MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems. "O'Reilly Media, Inc.", 2012. – 250 c.
3. [Thilina Gunarathne.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Thilina+Gunarathne%22&source=gbs_metadata_r&cad=7) Hadoop MapReduce v2 Cookbook – Second Edition. –: "Packt Publishing Ltd", 2015. – 322 c.
4. [Билл Фрэнкс.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22%D0%91%D0%B8%D0%BB%D0%BB+%D0%A4%D1%80%D1%8D%D0%BD%D0%BA%D1%81%22&source=gbs_metadata_r&cad=6) Укрощение больших данных: Как извлекать знания из массивов информации с помощью глубокой аналитики. –: "Манн, Иванов Фербер", 2014.
5. [Виктор Майер–Шенбергер](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22%D0%92%D0%B8%D0%BA%D1%82%D0%BE%D1%80+%D0%9C%D0%B0%D0%B9%D0%B5%D1%80-%D0%A8%D0%B5%D0%BD%D0%B1%D0%B5%D1%80%D0%B3%D0%B5%D1%80%22&source=gbs_metadata_r&cad=7), [Кеннет Кукьер](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22%D0%9A%D0%B5%D0%BD%D0%BD%D0%B5%D1%82+%D0%9A%D1%83%D0%BA%D1%8C%D0%B5%D1%80%22&source=gbs_metadata_r&cad=7). Большие данные: Революция,которая изменит то, как мы живем, работаем и мыслим. –: "Манн, Иванов, Фербер", 2013. – 240 c.
6. [Peter Bühlmann,](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Peter+B%C3%BChlmann%22&source=gbs_metadata_r&cad=7) [Petros Drineas,](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Petros+Drineas%22&source=gbs_metadata_r&cad=7) [Michael Kane,](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Michael+Kane%22&source=gbs_metadata_r&cad=7) [Mark van der Laan.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Mark+van+der+Laan%22&source=gbs_metadata_r&cad=7) Handbook of Big Data. –: "CRC Press", 2016. – 464 c.
7. [Frank J. Ohlhorst.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Frank+J.+Ohlhorst%22&source=gbs_metadata_r&cad=7) Big Data Analytics: Turning Big Data into Big Money. –: "John Wiley & Sons", 2012. – 176 c.
8. [Y. Lakshmi Prasad.](https://www.google.ru/search?hl=ru&tbo=p&tbm=bks&q=inauthor:%22Y.+Lakshmi+Prasad%22&source=gbs_metadata_r&cad=7) Big Data Analytics Made Easy. –: "Notion Press", 2016. -192 c.