Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования

«Московский государственный технический университет имени Н. Э. Баумана»

На правах рукописи

**Григорьев Ю.А.**

**Методические указания к курсовому проекту по дисциплине «Технология параллельных систем баз данных»**

**«Разработка макета аналитической системы на основе баз данных NoSQL»**

**Москва - 2024**

Оглавление

[1. Цель курсового проекта 3](#_Toc95052140)

[2. Задание 3](#_Toc95052141)

[3. Требования к оформлению курсового проекта 3](#_Toc95052142)

[3.1. Расчётно-пояснительная записка 3](#_Toc95052143)

[3.2. Перечень графического (иллюстративного) материала (плакатов) 4](#_Toc95052144)

[4. Литература 5](#_Toc95052145)

[5. Приложение 1. Титульный лист и задание (для РПЗ) 7](#_Toc95052146)

Внимание! Перед выполнением курсового проекта (КП) необходимо полностью ознакомиться с заданием и со своим вариантом проекта (варианты вышлю по почте).

# Цель курсового проекта

Разработать макет аналитической системы для указанного варианта предметной области с использованием баз данных NoSQL Elasticsearch, Neo4j, Hadoop+Spark.

# Задание

1. Установить виртуальную машину с ubuntu 20.04.3 (ubuntu-20.04.3-desktop-amd64.iso [1б]) в VirtualBox [1а] (python 3 уже предустановлен в ubuntu 20.04.3).

Имя: Ubuntu-01. Тип: Linux. Версия: Ubuntu (64 bit) и т.д. Объём ОП – 4ГБ, Диск – 20ГБ.

2. Установить Elasticsearch, Neo4j, Hadoop+Spark, как это Вы делали на лабораторных работах в 1-м семестре [2а]. Можете использовать виртуальное окружение Python [1с] (в будущем может очень пригодиться).

Примечание. Пункты 1 и 2 уже выполнены Вами в 1-м семестре при выполнении лабораторных работ по дисциплине ТПСБД

3. Решить следующие задачи (варианты будут разосланы по ЭП):

- вручную (или автоматически) создать два json-файла c 20÷30 json-документами каждого типа для предметной области, указанной в Вашем варианте; в варианте для каждого документа указаны его поля,

- в Elasticsearch: создать индекс с анализатором и маппингом [3], проиндексировать json-документы, разработать запросы с вложенной агрегацией [4], представить результаты в среде Kibana [5],

- в Neo4j: по данным из Elasticsearch заполнить графовую базу данных, разработать и реализовать запрос к этой БД,

- в Spark: по данным из Elasticsearch сформировать csv-файлы c таблицами и сохранить их в файловой системе HDFS, написать запрос и реализовать его в Spark, проанализировать процесс выполнения запроса с использованием монитора.

4. Дополнительные материалы будут выкладываться в папку «Курсовой проект 2024» на яндекс-диск [2б].

# Требования к оформлению курсового проекта

## Расчётно-пояснительная записка

РПЗ должна включать следующие разделы:

*Реферат.*

*Ведение.*

Здесь необходимо описать дать краткую характеристику Elasticsearch, Neo4j, Hadoop+Spark, описать цель КП.

*1. Задание и описание варианта КП.*

Привести задание и описать вариант КП.

*2. Elasticsearch.*

*2.1. Индексация документов.*

Текст описания анализатора и маппинга и его пояснение; алгоритм программы индексации документов и его пояснение.

*2.2. Запросы*.

Тексты запросов и их пояснение, результаты выполнения запросов.

*3. Neo4j.*

*3.1. Создание и заполнение графовой БД.*

Алгоритм программы создания и заполнения графовой БД и его пояснение.

3.2. *Запрос.*

Текст запроса на языке Cypher и его пояснение, результат выполнения запроса.

*4. Spark.*

*4.1. Создание и заполнение таблиц.*

Алгоритм программы создания csv-файлов с таблицами и их сохранения в HDFS, пояснение алгоритма.

*4.2. Запрос.*

Скрипт запроса к БД (с оператором select), пояснение, результат его выполнения.

*4.3. Монитор*.

Результат анализа работы монитора, пояснение.

*Заключение.*

*Список использованных источников.*

*Приложения:*

Привести распечатку одного json-документа каждого типа, тексты программ (**обязательно с подробными комментариями !!**), распечатку работы монитора Spark.

## Перечень графического (иллюстративного) материала (плакатов)

1. Название темы КП, задание, описание варианта.

2,3. По Elasticsearch: описание анализатора и маппинга; алгоритм программы индексации документов; тексты запросов, результаты выполнения.

4,5. По Neo4j: алгоритм программы создания и заполнения графовой БД; текст запроса на языке Cypher, результат выполнения запроса.

5,6,7. По Spark: алгоритм программы создания csv-файлов с таблицами и их сохранения в HDFS; скрипт запроса к БД (с оператором select), результат его выполнения; результат анализа работы монитора.

# Литература

1а. <https://white55.ru/vboxubuntu.html> установка ubuntu на ВМ

1б. <https://releases.ubuntu.com/20.04.3/> образ ubuntu-20.04.3-desktop-amd64.iso

1с. <https://pythonchik.ru/okruzhenie-i-pakety/virtualnoe-okruzhenie-python-venv> виртуальное окружение Python (venv)

2а. <https://disk.yandex.ru/d/41UnMXfWhT10yA> лаб.работы 2023

2б[.](.%20https:/disk.yandex.ru/d/lqltQcKL1bUpgg%20%20%20) <https://disk.yandex.ru/d/lqltQcKL1bUpgg> курсовой проект 2024

Для Elasticsearch:

3. Анализаторы и маппинг:

<https://habr.com/ru/post/280488/>

<https://xakep.ru/2015/06/11/elasticsearch-tutorial/#toc01>

<https://medium.com/nuances-of-programming/%D0%BD%D0%B0%D1%87%D0%B0%D0%BB%D0%BE-%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D1%8B-%D1%81-elasticsearch-%D0%B2-python-%D1%87%D0%B0%D1%81%D1%82%D1%8C-2-412796dcb213>

<https://kb.objectrocket.com/elasticsearch/how-to-map-an-elasticsearch-index-using-the-python-client-266>

4. Агрегация:

<https://habr.com/ru/company/mailru/blog/213849/>

<https://github.com/elastic/elasticsearch/issues/3300>

5. Kibana:

<https://coderlessons.com/tutorials/bolshie-dannye-i-analitika/vyuchit-kibanu/kibana-uchebnik>

6. Doc Elasticsearch 7.5:

<https://www.elastic.co/guide/en/elasticsearch/reference/7.5/index.html>

Для Neo4j:

7. py2neo:

<https://py2neo.org/2020.1/>

<https://py2neo.org/v4/index.html>

<https://question-it.com/questions/1689737/izmenit-ili-izmenit-svojstva-otnoshenij-v-py2neo>

8. cypher:

<https://neo4j.com/docs/cypher-manual/current/>

<http://art-in-stamps.ru/development/cypher-p45.shtml>

Для Spark

9. <https://spark.apache.org/docs/2.3.2/sql-programming-guide.html>

10. <https://ru.bmstu.wiki/index.php?title=Spark_SQL&mobileaction=toggle_view_mobile>

Дополнительная литература:

Для Elasticsearch:

<https://alexeykalina.github.io/technologies/elasticsearch-suggesters.html>

<https://alexeykalina.github.io/technologies/elasticsearch-autocomplete.html>

<https://alexeykalina.github.io/technologies/elasticsearch-fulltextsearch.html>

<https://alexeykalina.github.io/technologies/elasticsearh-facets.html>

ML с Elasticsearch:

<https://habr.com/ru/company/galssoftware/blog/455387/>

<https://medium.com/@bigdataschool/%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D1%82%D0%B8%D0%BA%D0%B0-%D0%B1%D0%BE%D0%BB%D1%8C%D1%88%D0%B8%D1%85-%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85-%D0%B2-elasticsearch-%D0%B2%D0%BE%D0%B7%D0%BC%D0%BE%D0%B6%D0%BD%D0%BE%D1%81%D1%82%D0%B8-machine-learning-%D0%B2-elk-stack-6d5e4e6e6dd0>

Для Neo4j

<https://habr.com/ru/post/219441/>

Для Spark

<https://www.machinelearningmastery.ru/working-with-apache-spark-python-and-pyspark-128a82668e67/>

<https://github.com/big-data-europe/docker-hadoop-spark-workbench>

<https://github.com/big-data-europe/docker-spark>

# Приложение 1. Титульный лист и задание (для РПЗ)

|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ Информатика и системы управления

КАФЕДРА ИУ6

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К КУРСОВОМУ ПРОЕКТУ***

***НА ТЕМУ:***

**Разработка макета аналитической системы на основе баз данных NoSQL (вариант № )**

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

Студент ИУ6-2хМ\_\_\_\_\_\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель курсового проекта **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Григорьев Ю.А.

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2024 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_\_\_\_\_\_\_\_

(Индекс)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(И.О.Фамилия)

« \_\_\_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20 \_\_\_\_ г.

**ЗАДАНИЕ**

**на выполнение курсового проекта**

по дисциплине «Технология параллельных систем баз данных»

Студент группы ИУ6-23М

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Тема курсового проекта

«Разработка макета аналитической системы на основе баз данных NoSQL (вариант № )»

Направленность КП (учебный, исследовательский, практический, производственный, др.)

исследовательский

Источник тематики (кафедра, предприятие, НИР) кафедра

График выполнения проекта: 25% к 2 нед., 50% к 10 нед., 75% к 13 нед., 100% к 16 нед.

***Задание***

Установить виртуальную машину, системы NoSQL Elasticsearch, Neo4j, Hadoop+Spark (выполнено на 1-м семестре). В **Elasticsearch** создать индекс с анализатором и маппингом, проиндексировать json-документы, разработать запросы с вложенной агрегацией, представить результаты в среде Kibana. В **Neo4j** по данным из Elasticsearch заполнить графовую базу данных, разработать и реализовать запрос к этой БД. В **Spark** по данным из Elasticsearch сформировать csv-файлы (с внутренней схемой) таблиц и сохранить их в файловой системе HDFS, написать запрос и реализовать его в Spark, проанализировать процесс выполнения запроса с использованием монитора.

***Оформление курсового проекта:***

Расчетно-пояснительная записка на \_\_\_ листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.):

1. Название темы КП, задание, описание варианта.

2,3. По Elasticsearch: описание анализатора и маппинга; алгоритм программы индексации документов; тексты запросов, результаты выполнения.

4,5. По Neo4j: алгоритм программы создания и заполнения графовой БД; текст запроса, результат.

5,6,7. По Spark: алгоритм программы создания таблиц и их сохранения в HDFS; скрипт запроса к БД, результат выполнения; результат анализа работы монитора.

Дата выдачи задания « \_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20\_\_ г.

**Руководитель курсового проекта**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_ Григорьев Ю.А.

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.