Разработка макета аналитической системы на основе баз данных NoSQL (вариант № 6)

Задание:

1) Установить виртуальную машину с ОС Ubuntu в VirtualBox.
2) Установить Elasticsearch, Neo4j, Hadoop, Spark.
3) Вручную создать JSON-файл с 20-30 JSON-документами для предметной области, указанной в варианте.
4) В Elasticsearch создать индекс с анализатором и маппингом, проиндексировать JSON-документы, разработать запросы с вложенной агрегацией, представить результаты в среде Kibana.
5) В Neo4j по данным из Elasticsearch заполнить графовую базу данных, разработать и реализовать запрос к этой БД.
6) В Spark по данным из Elasticsearch сформировать csv-файлы с таблицами и сохранить их в файловой системе HDFS, написать запрос и реализовать его в Spark, проанализировать процесс выполнения запроса с использованием монитора Spark.

Предметная область – Салон красоты

Elasticsearch:

1. Типы JSON-документов:

Процедура:
{index, doc_type, id, body: {id_клиента, возраст, персональные_данные*, id_процедуры, дата_процедуры, стоимость, id_специалиста, [услуга*], [препарат*]}}
Специалист:
{index, doc_type, id, body: {специализация, стаж_работы, сведения_о_специалисте, график_работы*, [отзыв_о_специалисте*] }}

2. Требования к анализатору: для полей, отмеченных символом *, должен быть разработан анализатор со следующими требованиями: разделить текст на слова, убрать пунктуацию с помощью токенизатора standard (русский), перевести все токены в нижний регистр, убрать токены, находящиеся в списке стоп-слов, выполнить стемминг оставшихся токенов с помощью фильтра snowball.
3. Запросы с вложенной агрегацией:
   – разбить процедуры по дате с периодом 1 год, для каждой "корзины" определить суммарную стоимость по каждому специалисту,
   – предложить признаки отрицательного отзыва; вывести специалистов хотя бы с одним отрицательным отзывом.

Neo4j:

1. По данным из Elasticsearch заполнить графовую базу данных Клиент(id_клиента, персональные_данные) - Посетил(дата_процедуры, стоимость) - Специалист(id_специалиста, специализация, стаж работы).
2. Разработать и реализовать запрос: определить специалиста с наибольшей стоимостью выполненных процедур.

Spark:

1. По данным из Elasticsearch сформировать csv-файлы (с внутренней схемой) таблиц "Клиент", "Процедура", "Специалист" и сохранить их в файловой системе HDFS.
2. Написать запрос select: определить число выполненных процедур каждым специалистом.
3. Реализовать этот запрос в Spark. Построить временную диаграмму его выполнения по результатам работы монитора.

					Разработка макета аналитической системы на основе баз данных NoSQL (вариант № 6)	
Изм	Лист	№ документа	Подпись	Дата	Задание на курсовой проект	Литер. / Масса / Масштаб
Разраб.		Матвиенко Е.К.				Лист 1 / Листов 11
Руковод.		Григорьев Ю. А.				
Н. Контр.						МГТУ им. Н. Э. Баумана Группа ИУ6-23М

Индексация документов Elasticsearch

Анализатор для индексов

```
"analysis" : {
    "filter": {
        "russian_stop_words": {
            "type": "stop",
            "stopwords": "_russian_"
        },
        "filter_ru_sn": {
            "type": "snowball",
            "language":"Russian"
        }
    },
    "analyzer":
    {
        "analitic_for_ru":
        {
            "type": "custom",
            "tokenizer": "standard",
            "filter": [
                "lowercase",
                "russian_stop_words",
                "filter_ru_sn"
            ]
        }
    }
}
```

Фрагмент маппинга для индекса процедура

```
ProcedureMapping = {
    "properties":{
        "id_of_client": {
            "type": "text",
            "fielddata": True
        },
        "client_age": {
            "type": "integer"
        },
        "client_personal_data":{
            "type": "text",

"analyzer":"analitic_for_ru",
            "fielddata": True
        },
        "date_of_procedure": {
            "type":   "date",
            "format": "yyyy-MM-dd"
        }
...
    }
}
```

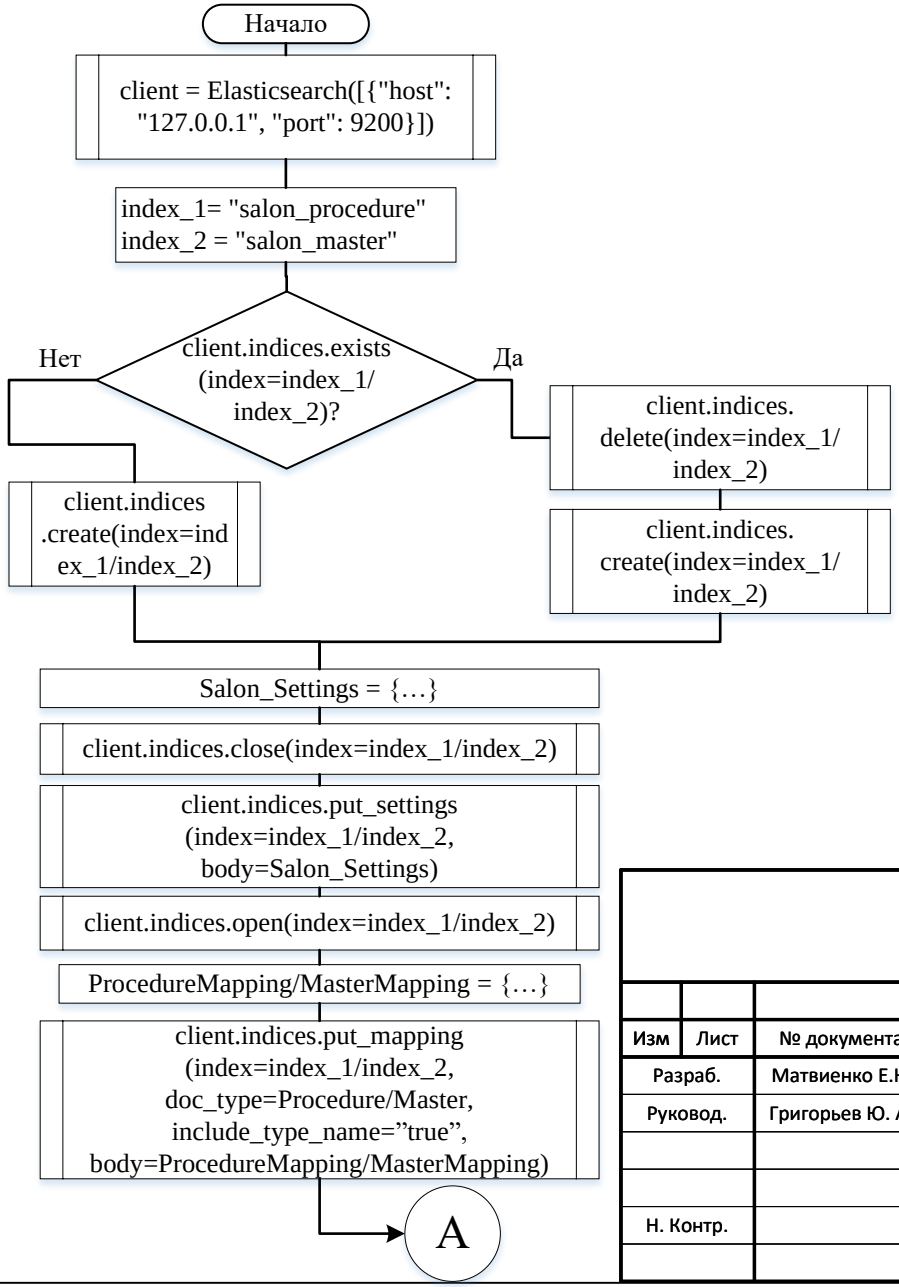
Фрагмент маппинга для индекса мастер

```
MasterMapping = {
    "properties":{
        "specialisation": {
            "type": "text",
            "fielddata": True,
            "analyzer":"analitic_for_ru",

"search_analyzer":"analitic_for_ru"

        },
...
         "reviews": {
            "type": "text",
            "fielddata": True,
            "analyzer":"analitic_for_ru",

"search_analyzer":"analitic_for_ru"
        }
    }
}
```

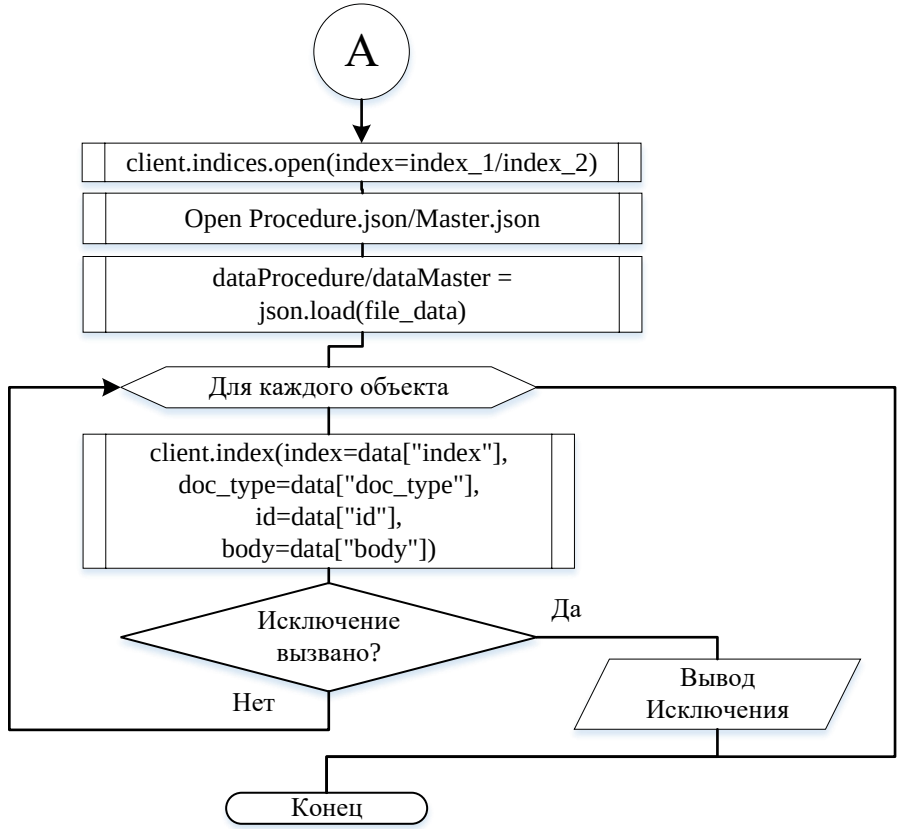
				Разработка макета аналитической системы на основе баз данных NoSQL (вариант № 6)
Изм	Лист	№ документа	Подпись	Дата	Индексация документов Elasticsearch. Маппинг и анализатор	Литер.	Масса	Масштаб
Разраб.		Матвиенко Е.К.						
Руковод.		Григорьев Ю. А.				Лист 2	Листов 11	
Н. Контр.						МГТУ им. Н. Э. Баумана Группа ИУ6-23М		

Индексация документов Elasticsearch

Алгоритм добавления маппинга и настройки анализатора



Алгоритм индексации



Изм	Лист	№ документа	Подпись	Дата
Разраб.		Матвиенко Е.К.		
Руковод.		Григорьев Ю. А.		
Н. Контр.				

Разработка макета аналитической системы на основе баз данных NoSQL (вариант № 6)

Индексация документов Elasticsearch. Алгоритмы индексациии

Литер.	Масса	Масштаб

Лист 3 | Листов 11

МГТУ им. Н. Э. Баумана
Группа ИУ6-23М

Elasticsearch. Запросы

Первый запрос: разбить процедуры по дате с периодом 1 год, для каждой "корзины" определить суммарную стоимость по каждому специалисту

```
GET salon_procedure/_search
{
  "size": 0,
  "aggregations": {
    "year_period": {
      "date_histogram": {
        "field": "date_of_procedure",
        "calendar_interval": "year",
        "format": "yyyy-MM-dd"
      },
      "aggregations": {
        "Specialist": {
          "terms": {
            "field": "id_of_specialist",
            "order": {
              "_key": "asc"
            }
          },
          "aggregations": {
            "summary_price": {
              "sum": {
                "field": "price"
              }
            }
          }
        }
      }
    }
  }
}
```

<details>
<summary>Ответ</summary>

```
{
  "took" : 90,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 40,
      "relation" : "eq"
    },
    "max_score" : null,
    "hits" : [ ]
  },
  "aggregations" : {
    "year_period" : {
      "buckets" : [
        {
          "key_as_string" : "2016-01-01",
          "key" : 1451606400000,
          "doc_count" : 7,
          "Specialist" : {
            "doc_count_error_upper_bound" : 0,
            "sum_other_doc_count" : 0,
            "buckets" : [
              {
                "key" : "sp04",
                "doc_count" : 1,
                "summary_price" : {
                  "value" : 6418.0
                }
              },
              {
                "key" : "sp06",
                "doc_count" : 1,
                "summary_price" : {
                  "value" : 5559.0
                }
              },
              {
                "key" : "sp07",
                "doc_count" : 1,
                "summary_price" : {
                  "value" : 4988.0
                }
              },
```

</details>

Второй запрос: предложить признаки отрицательного отзыва; вывести специалистов хотя бы с одним отрицательным отзывом

```
GET salon_master/_search
{
  "query": {
    "simple_query_string": {
      "query": "кидает пострадало расстроенный наглость опоздал помойка грязно ужасно плохо печально"
    }
  },
  "_source": [
    "master_personal_data"
  ],
  "highlight": {
    "fields": {
      "reviews": {}
    }
  }
}
```

<details>
<summary>Ответ</summary>

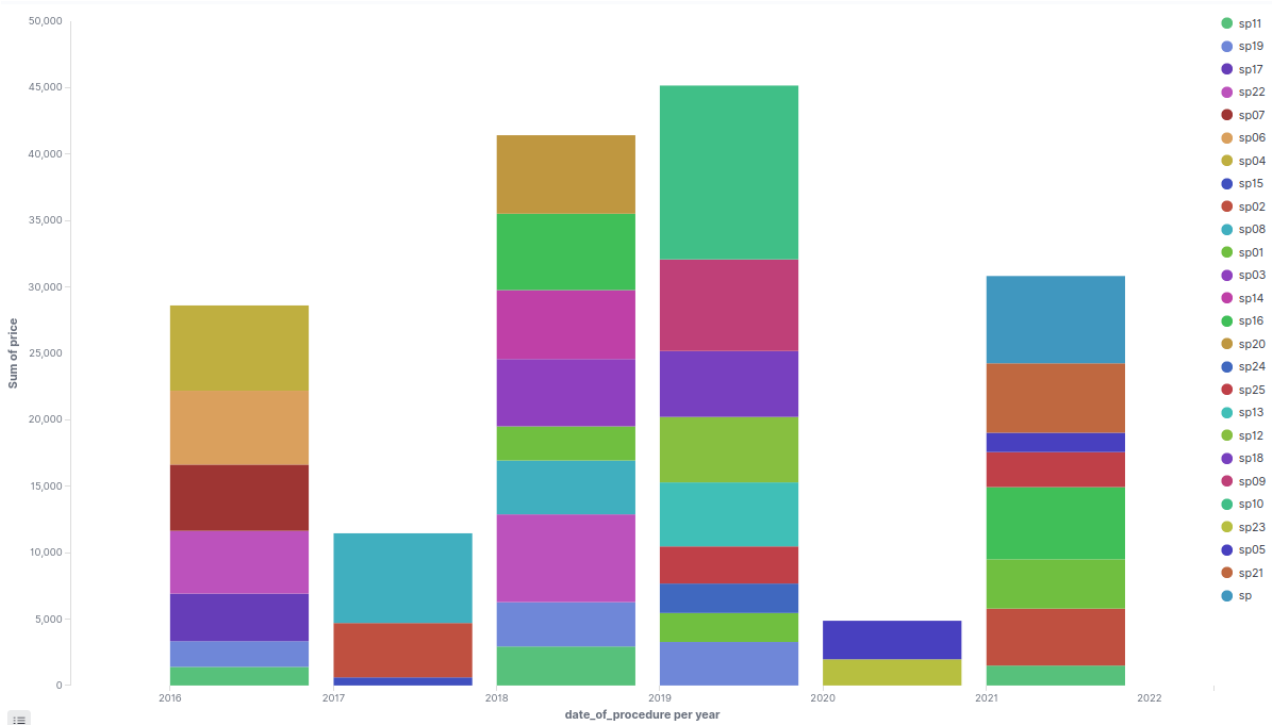
```
{
  "took" : 243,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 14,
      "relation" : "eq"
    },
    "max_score" : 4.650751,
    "hits" : [
      {
        "_index" : "salon_master",
        "_type" : "Master",
        "_id" : "SP04",
        "_score" : 4.650751,
        "_source" : {
          "master_personal_data" : "Семина Анна Игоревна"
        },
        "highlight" : {
          "reviews" : [
            "Первый раз в жизни пишу негативный отзыв, но тут уже нет предела <em>наглости</em>, а я наглых людей не люблю.",
            "Если вы читаете этот отзыв, то знайте, что мастер <em>плохо</em> подходит к своему делу.",
            "Работа выполнена <em>ужасно</em>."
          ]
        }
      }
```

</details>

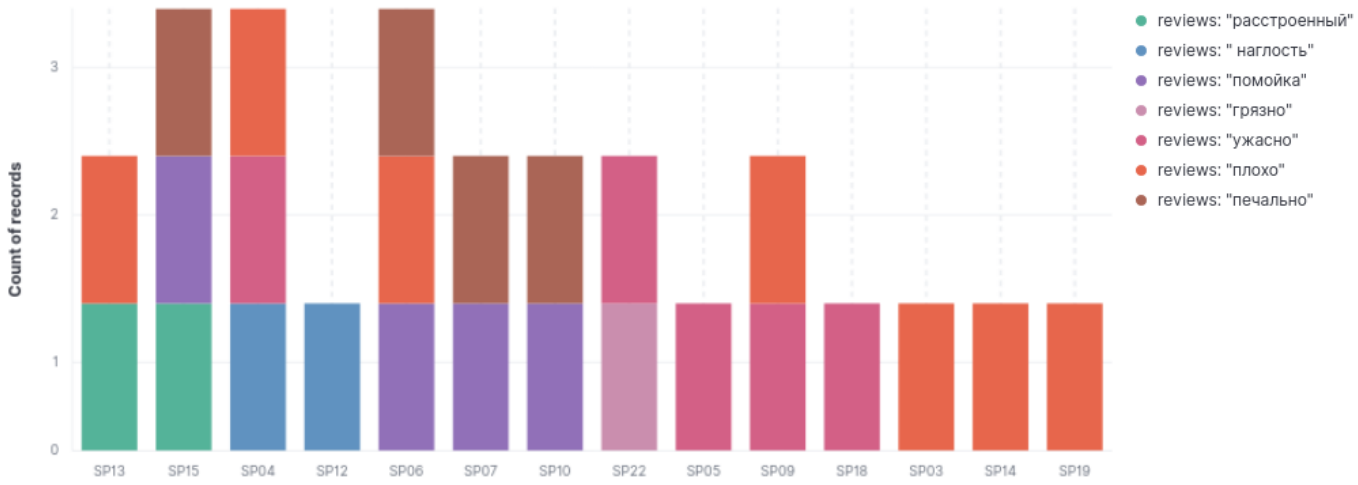
					Разработка макета аналитической системы на основе баз данных NoSQL (вариант № 6)	
Изм	Лист	№ документа	Подпись	Дата	Elasticsearch. Запросы	Литер. / Масса / Масштаб
Разраб.		Матвиенко Е.К.				Лист 4 / Листов 11
Руковод.		Григорьев Ю. А.				МГТУ им. Н. Э. Баумана Группа ИУ6-23М
Н. Контр.						

Elasticsearch-Kibana. Визуализация запросов

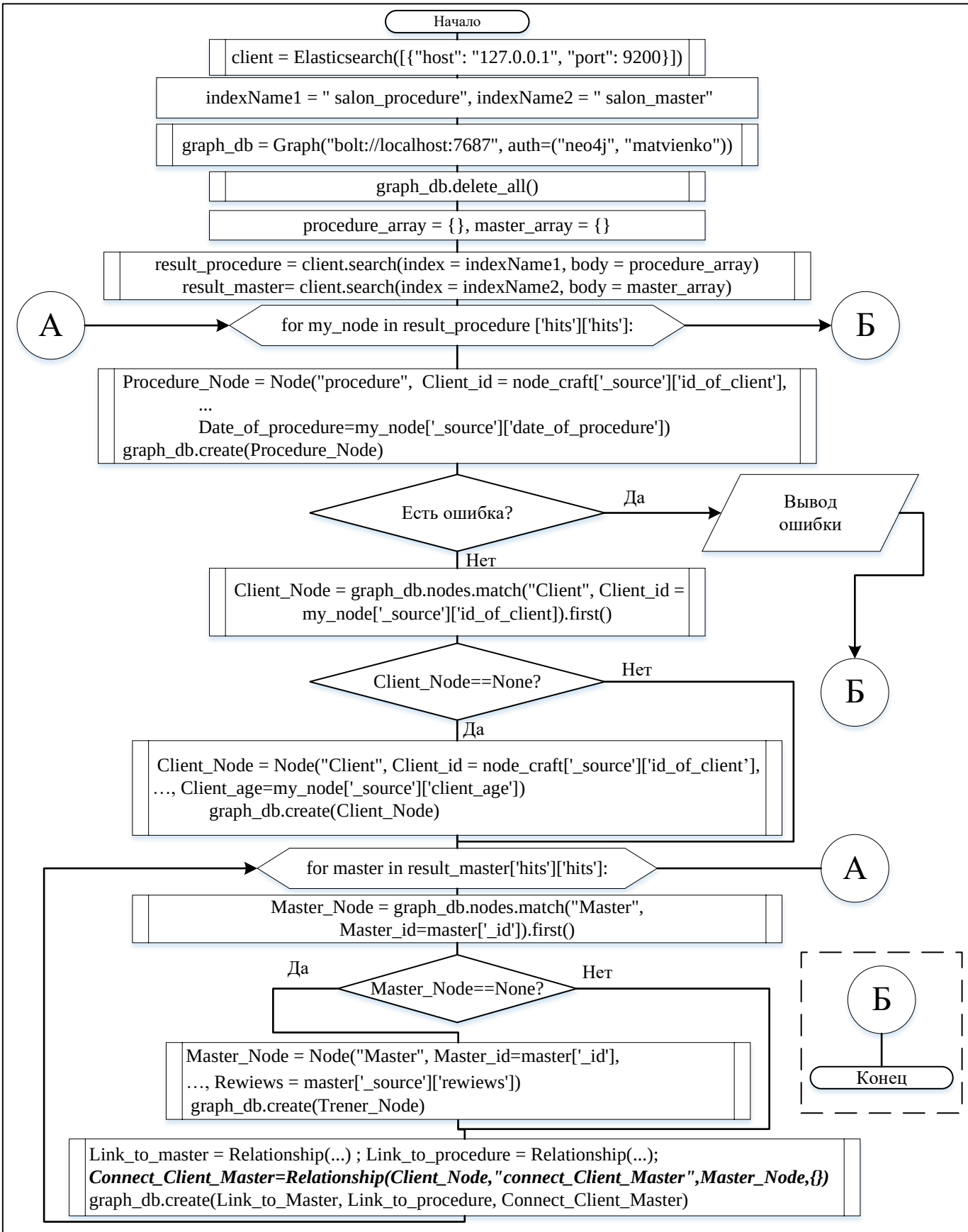
Визуализация первого запроса



Визуализация второго запроса



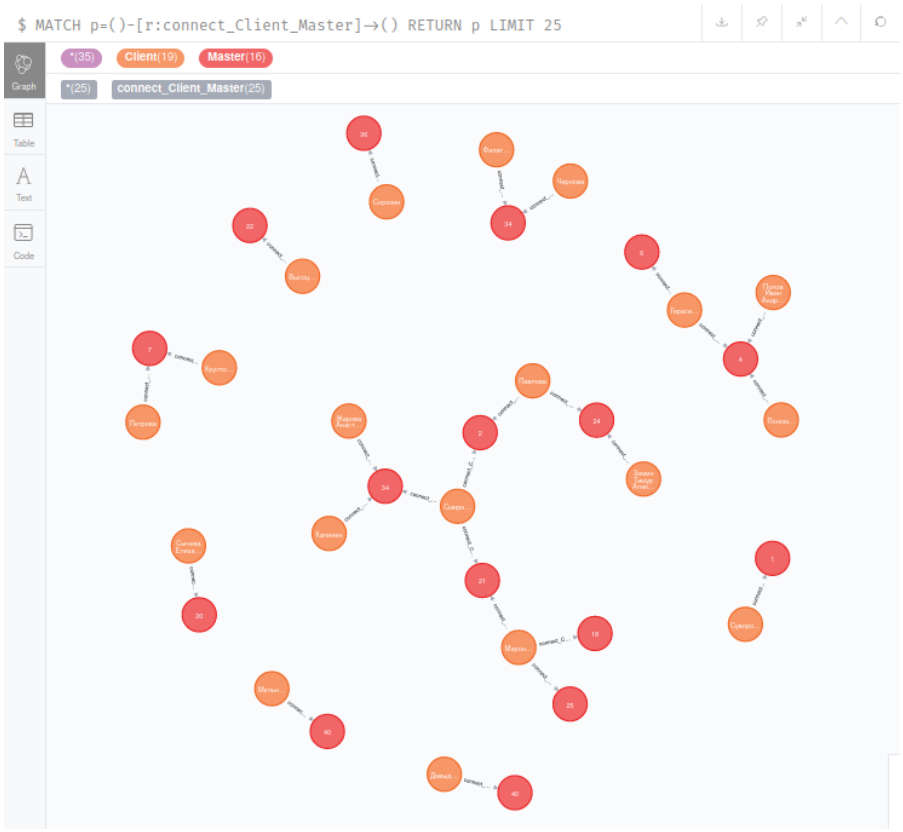
					Разработка макета аналитической системы на основе баз данных NoSQL (вариант № 6)	
					Elasticsearch. Визуализация запросов	Литер. / Масса / Масштаб
Изм	Лист	№ документа	Подпись	Дата		
Разраб.		Матвиенко Е.К.				Лист 5 / Листов 11
Руковод.		Григорьев Ю. А.				
Н. Контр.						МГТУ им. Н. Э. Баумана Группа ИУ6-23М



					Разработка макета аналитической системы на основе баз данных NoSQL (вариант № 6)			
						Литер.	Масса	Масштаб
Изм	Лист	№ документа	Подпись	Дата	Neo4j. Алгоритм создания и заполнения графовой БД			
Разраб.		Матвиенко Е.К.				Лист 6	Листов 11	
Руковод.		Григорьев Ю. А.						
Н. Контр.						МГТУ им. Н. Э. Баумана Группа ИУ6-23М		

Neo4j. Запрос

Связь «клиент посетил специалиста» = = Connect_Client_Master



Запрос: определить специалиста с наибольшей стоимостью выполненных процедур

```
MATCH p=(c:Client)-[r:connect_Client_Master]→(m:Master)
WITH m, sum(toInteger(r.Price)) as master_sum
ORDER BY master_sum desc
RETURN m, master_sum
LIMIT 1
```

$ MATCH p=(c:Client)-[r:connect_Client_Master]→(m:Master) WITH m…

m	master_sum
	13084


```
{
  "Work_experience": "2",
  "Master_id": "SP10",
  "Specialization": [
    "Массаж рук",
    "Spa-маникюр",
    "Биопилинг кожи головы",
    "Гальванизация"
  ],
  "Master_personal_data":
"Ильинский Матвей Эмирович"
}
```



```
{
  "Work_experience": "40",
  "Master_id": "SP04",
  "Specialization": [
    "Маникюр Европейский",
    "Уход за руками",
    "Долговременная укладка",
    "Долгосрочное покрытие гель-
лаками (ши-лаками)"
  ],
  "Master_personal_data": "Семина
Анна Игоревна"
}
```



```
{

"Client_personal_data":
"Давыдова Маргарита
Матвеевна",
  "Client_age": "20",
  "Client_id": "HXAB"
}
```


Разработка макета аналитической системы на основе баз данных NoSQL (вариант № 6)

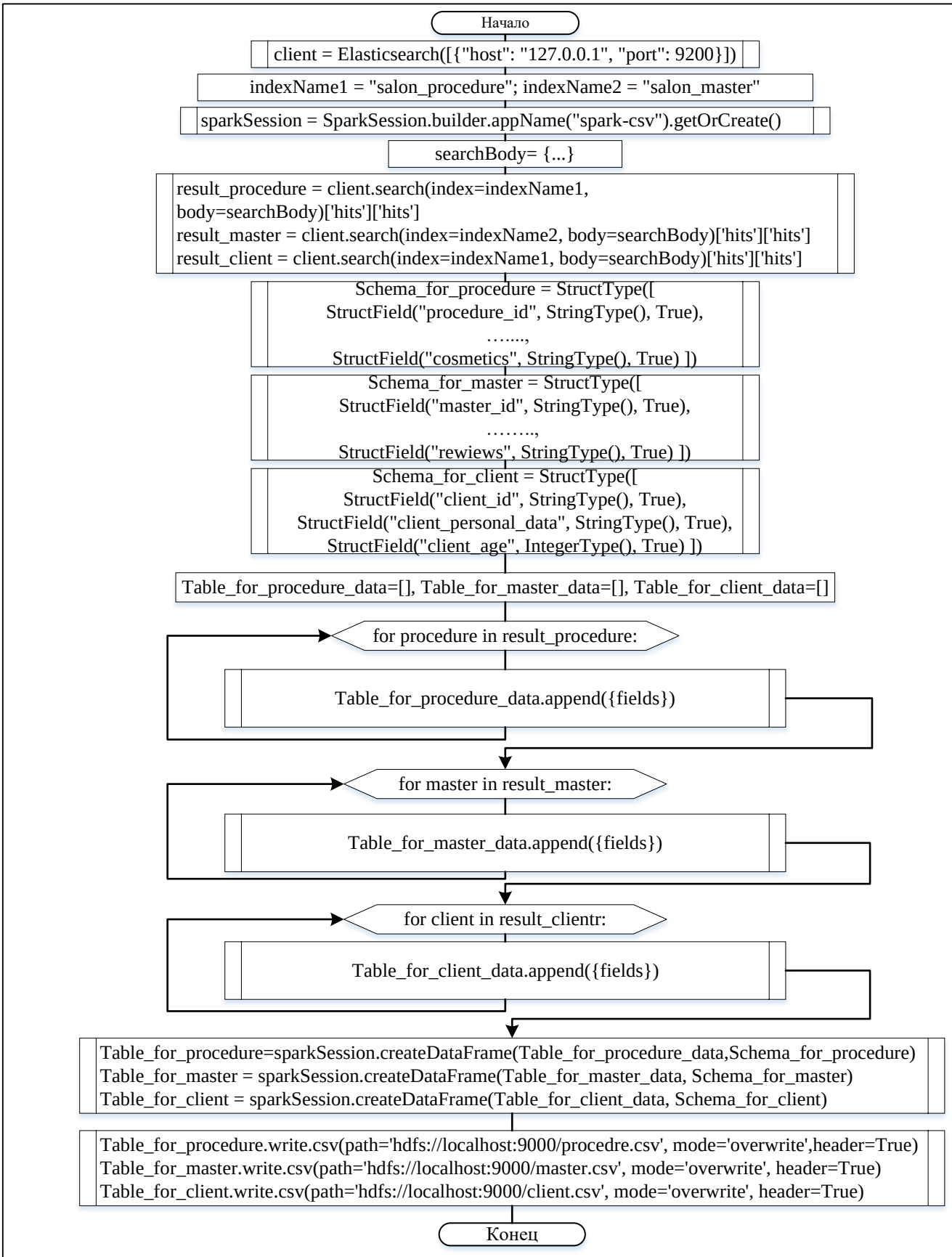
Изм	Лист	№ документа	Подпись	Дата
Разраб.		Матвиенко Е.К.		
Руковод.		Григорьев Ю. А.		
Н. Контр.				

Neo4j. Запрос

Литер.	Масса	Масштаб

Лист 7 | Листов 11

МГТУ им. Н. Э. Баумана
Группа ИУ6-23М



					Разработка макета аналитической системы на основе баз данных NoSQL (вариант № 6)	
Изм	Лист	№ документа	Подпись	Дата	Spark. Алгоритм создания CSV-файлов с таблицами	Литер. / Масса / Масштаб
Разраб.		Матвиенко Е.К.				
Руковод.		Григорьев Ю. А.				Лист 8 / Листов 11
Н. Контр.						МГТУ им. Н. Э. Баумана Группа ИУ6-23М

Spark. Запрос

Запрос: определить число выполненных процедур каждым специалистом

```
+---------+--------------------+
|master_id|count(procedure_id)|
+---------+--------------------+
|     SP05|                   2|
|     SP19|                   3|
|     SP14|                   1|
|     SP20|                   1|
|     SP18|                   1|
|     SP24|                   2|
|     SP16|                   2|
|     SP09|                   1|
|     SP01|                   3|
|     SP04|                   1|
|       SP|                   1|
|     SP02|                   2|
|     SP22|                   2|
|     SP10|                   2|
|     SP11|                   3|
|     SP21|                   1|
|     SP12|                   1|
|     SP23|                   1|
|     SP17|                   1|
|     SP25|                   2|
|     SP13|                   1|
|     SP15|                   1|
|     SP07|                   1|
|     SP06|                   1|
|     SP08|                   2|
|     SP03|                   1|
+---------+--------------------+
```

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-from pyspark.sql import SparkSession

from pyspark.sql import *

sparkSession=SparkSession.builder.appName("Python Spark SQL basic
example").config("spark.sql.shuffle.partitions","10").getOrCreate()

Procedure_Table = sparkSession.read.load(path='hdfs://localhost:9000/procedure.csv', format='csv', sep=',',
        inferSchema="true", header="true")
Master_Table = sparkSession.read.load(path='hdfs://localhost:9000/master.csv', format='csv', sep=',',
        inferSchema="true", header="true")
Client_Table = sparkSession.read.load(path='hdfs://localhost:9000/client.csv', format='csv', sep=',',
        inferSchema="true", header="true")

Procedure_Table.registerTempTable("procedure")
Master_Table.registerTempTable("master")
Client_Table.registerTempTable("client")

df = sparkSession.sql("SELECT DISTINCT procedure.master_id, COUNT(procedure.procedure_id)
FROM procedure GROUP BY procedure.master_id").show(26)

input('Ctrl C')
```

					Разработка макета аналитической системы на основе баз данных NoSQL (вариант № 6)	
Изм	Лист	№ документа	Подпись	Дата	Spark. Запрос	Литер. / Масса / Масштаб
Разраб.		Матвиенко Е.К.				
Руковод.		Григорьев Ю. А.				Лист 9 / Листов 11
Н. Контр.					МГТУ им. Н. Э. Баумана Группа ИУ6-23М	

Spark. Мониторинг выполнения запроса

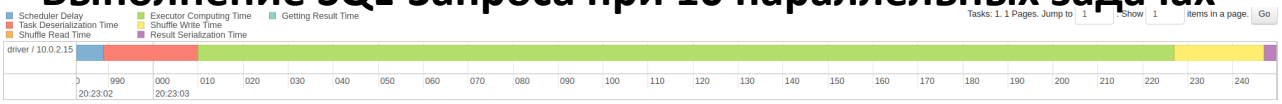
Выполненные SQL-запросы

ID ▾	Description	Submitted	Duration	Job IDs
6	showString at NativeMethodAccessorImpl.java:0 +details	2022/05/24 20:23:02	0,9 s	[6][7]
5	createOrReplaceTempView at NativeMethodAccessorImpl.java:0 +details	2022/05/24 20:23:02	4 ms	
4	createOrReplaceTempView at NativeMethodAccessorImpl.java:0 +details	2022/05/24 20:23:02	1 ms	
3	createOrReplaceTempView at NativeMethodAccessorImpl.java:0 +details	2022/05/24 20:23:02	22 ms	
2	load at NativeMethodAccessorImpl.java:0 +details	2022/05/24 20:23:02	84 ms	[4]
1	load at NativeMethodAccessorImpl.java:0 +details	2022/05/24 20:23:01	0,1 s	[2]
0	load at NativeMethodAccessorImpl.java:0 +details	2022/05/24 20:22:57	2 s	[0]

Spark Jobs

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
7	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2022/05/24 20:23:03	0,1 s	1/1 (1 skipped)	1/1 (1 skipped)
6	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2022/05/24 20:23:02	0,3 s	1/1	1/1
5	load at NativeMethodAccessorImpl.java:0 load at NativeMethodAccessorImpl.java:0	2022/05/24 20:23:02	41 ms	1/1	1/1
4	load at NativeMethodAccessorImpl.java:0 load at NativeMethodAccessorImpl.java:0	2022/05/24 20:23:02	39 ms	1/1	1/1
3	load at NativeMethodAccessorImpl.java:0 load at NativeMethodAccessorImpl.java:0	2022/05/24 20:23:01	96 ms	1/1	1/1
2	load at NativeMethodAccessorImpl.java:0 load at NativeMethodAccessorImpl.java:0	2022/05/24 20:23:01	45 ms	1/1	1/1
1	load at NativeMethodAccessorImpl.java:0 load at NativeMethodAccessorImpl.java:0	2022/05/24 20:23:00	0,9 s	1/1	1/1
0	load at NativeMethodAccessorImpl.java:0 load at NativeMethodAccessorImpl.java:0	2022/05/24 20:22:58	1 s	1/1	1/1

Выполнение SQL-Запроса при 10 параллельных задачах



Summary Metrics for 1 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	0.2 s	0.2 s	0.2 s	0.2 s	0.2 s
GC Time	0.0 ms	0.0 ms	0.0 ms	0.0 ms	0.0 ms
Input Size / Records	8.3 KiB / 40	8.3 KiB / 40	8.3 KiB / 40	8.3 KiB / 40	8.3 KiB / 40
Shuffle Write Size / Records	876 B / 26	876 B / 26	876 B / 26	876 B / 26	876 B / 26

Showing 1 to 4 of 4 entries

Aggregated Metrics by Executor

Executor ID	Logs	Address	Task Time	Total Tasks	Failed Tasks	Killed Tasks	Succeeded Tasks	Excluded	Input Size / Records	Shuffle Write Size / Records
driver		10.0.2.15:35497	0.3 s	1	0	0	1	false	8.3 KiB / 40	876 B / 26

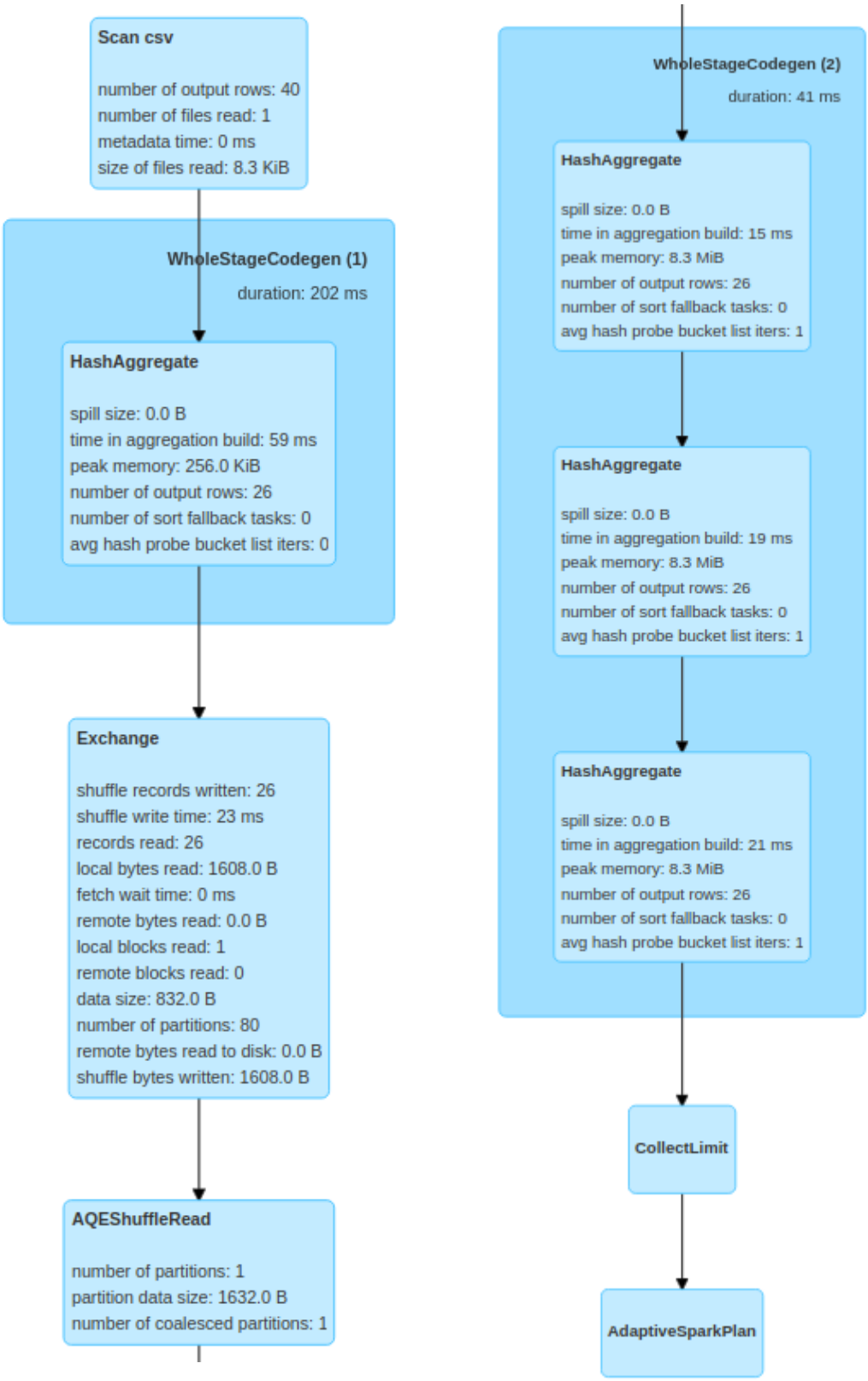
Showing 1 to 1 of 1 entries

Tasks (1)

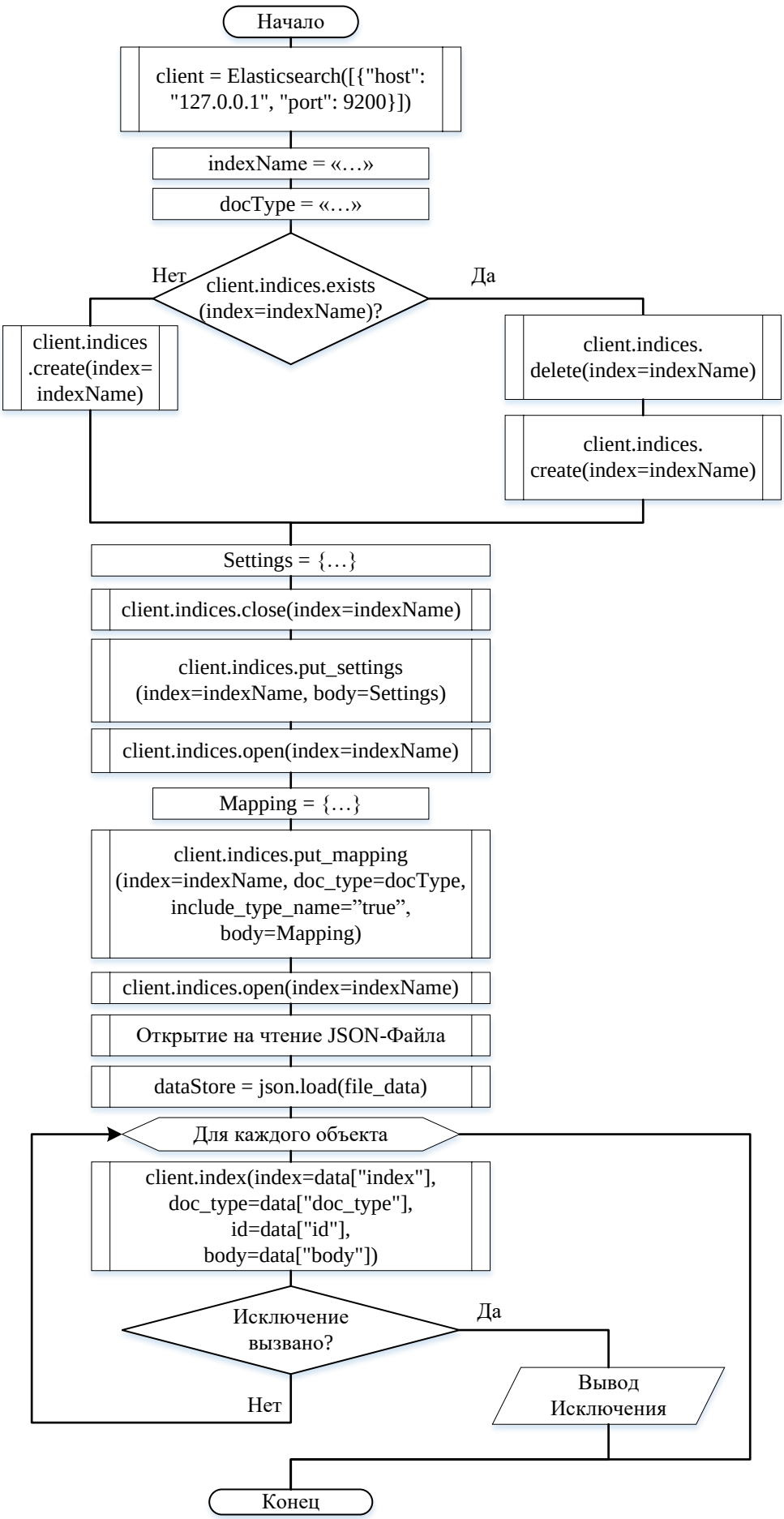
Index	Task ID	Attempt	Status	Locality level	Executor ID	Host	Logs	Launch Time	Duration	GC Time	Input Size / Records	Shuffle Write Size / Records	Errors
0	6	0	SUCCESS	ANY	driver	10.0.2.15		2022-05-24 20:23:02	0.2 s		8.3 KiB / 40	876 B / 26	

					Разработка макета аналитической системы на основе баз данных NoSQL (вариант № 6)			
Изм	Лист	№ документа	Подпись	Дата	Spark. Мониторинг выполнения запроса	Литер.	Масса	Масштаб
Разраб.		Матвиенко Е.К.						
Руковод.		Григорьев Ю. А.				Лист 10	Листов 11	
Н. Контр.						МГТУ им. Н. Э. Баумана Группа ИУ6-23М		

Spark. DAG SQL-запроса



					Разработка макета аналитической системы на основе баз данных NoSQL (вариант № 6)	
Изм	Лист	№ документа	Подпись	Дата	Spark. DAG SQL-запроса	Литер. / Масса / Масштаб
Разраб.		Матвиенко Е.К.				
Руковод.		Григорьев Ю. А.				Лист 11 / Листов 11
Н. Контр.						МГТУ им. Н. Э. Баумана Группа ИУ6-23М


Начало
client = Elasticsearch([{"host": "127.0.0.1", "port": 9200}])
indexName = «…»
docType = «…»
client.indices.exists (index=indexName)?
Нет
Да
client.indices .create(index= indexName)
client.indices. delete(index=indexName)
client.indices. create(index=indexName)
Settings = {…}
client.indices.close(index=indexName)
client.indices.put_settings (index=indexName, body=Settings)
client.indices.open(index=indexName)
Mapping = {…}
client.indices.put_mapping (index=indexName, doc_type=docType, include_type_name=”true”, body=Mapping)
client.indices.open(index=indexName)
Открытие на чтение JSON-Файла
dataStore = json.load(file_data)
Для каждого объекта
client.index(index=data["index"], doc_type=data["doc_type"], id=data["id"], body=data["body"])
Исключение вызвано?
Да
Нет
Вывод Исключения
Конец