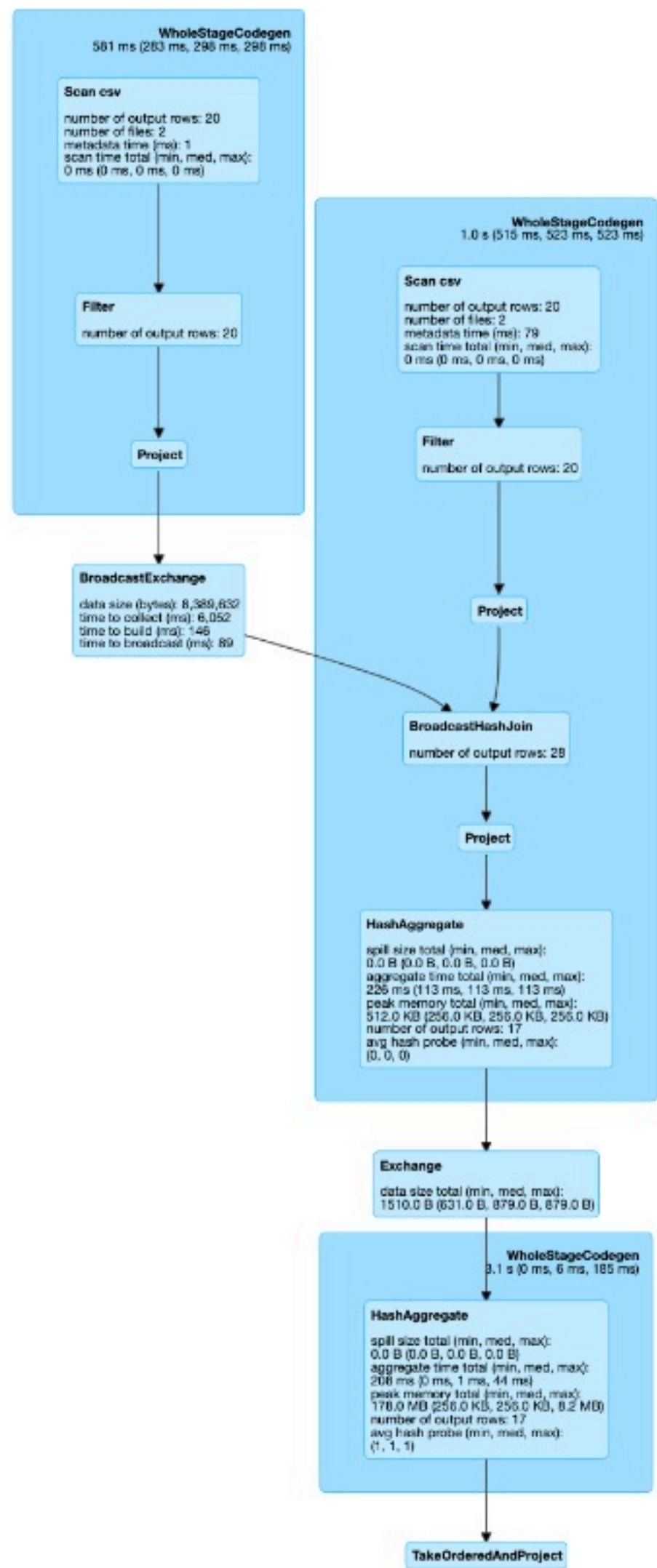


DAG SQL-Запроса



					Разработка макета аналитической системы строительная компания											
					Spark. DAG SQL-Запроса						Лит.		Масса	Масштаб		
Изм	Лист	№ докум.	Подп.	Дата							Лист 11		Листов			
Разраб.		Горский Г.Е.														
Проверил		Григорьев Ю.А.														
											МГТУ им. Баумана					
											ИУ6-23М					

# Разработка макета аналитической системы на основе баз данных NoSQL (вариант № 5)

## Задание

- 1 Установить виртуальную машину с ubuntu 20.04.3 в VirtualBox
- 2 Установить Elasticsearch, Neo4j, Hadoop+Spark.
- 3 Решить следующие задачи:
- создать два json-файла с 20–30 json-документами каждого типа для предметной области, указанной в варианте;

– в Elasticsearch: создать индекс с анализатором и маппингом, проработать 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 таблицами и сохранить их в файловой системе HDFS, написать запрос и реализовать его в Spark, проработать процесс выполнения запроса с использованием монитора.

## Вариант

Elasticsearch.

1. Типы документов (json):

Заказ:

```
{index: doc_type, id, body: {id_заказа, дата_заказа, id_заказчика, сведения_о_заказе *, данные_о_заказе *, срок_выполнения_заказа, фактическая_дата_выполнения, стоимость_заказа, id_бригады}}
```

Бригада:

```
{index: doc_type, id, body: {сведения_о_бригаде *, {член_бригады}, {отзыв_о_работе*}}}
```

2. Требования к анализатору:

поля, отмеченные \*, разделить 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 заказы по дате заказа с периодом 1 год, для каждой группы определить суммарную стоимость заказов по каждой бригаде;

– предложить признаки отрицательного отзыва; определить бригады хотя бы с одним отрицательным отзывом.

Neo4j.

1. По данным из Elasticsearch заполнить графовую базу данных:

Заказ(id\_заказа, дата\_заказа, сведения\_о\_заказчике, стоимость\_заказа)–Выполнил(срок\_выполнения\_заказа, фактическая\_дата\_выполнения) – Бригада(id\_бригады, сведения\_о\_бригаде).

2. Разработать и реализовать запрос: найти заказы и бригады, которые выполнили заказы с превышением срока

Spark

1. По данным из Elasticsearch сформировать csv-файлы (с внутренней схемой) таблиц «Заказчик», «Заказ», «Бригада» и сохранить их в файловой системе HDFS.

2. Написать запрос select: найти суммарную стоимость заказов по каждому заказчику.

3. Реализовать этот запрос в Spark. Построить временную диаграмму его выполнения по результатам работы монитора.

				Разработка макета аналитической системы			
				Строительная компания			
				Задание на курсовой проект			
Изм		Лист	№ докум.	Подпись	Дата		
Разраб.			Горский Г.Е.				
Провер.			Григорьев Ю.А.				

Малпунг "Заказ"

```
mapping = {
  "properties": {
    "id_заказа": {
      "type": "integer"
    },
    "дата_заказа": {
      "type": "date",
      "format": "YYYY-mm-dd"
    },
    "id_заказчика": {
      "type": "integer"
    },
    "сведения_о_заказчике": {
      "type": "text",
      "analyzer": "my_custom_analyzer"
    },
    "данные_о_заказе": {
      "type": "text",
      "analyzer": "my_custom_analyzer"
    },
    "срок_выполнения_заказа": {
      "type": "date",
      "format": "YYYY-mm-dd"
    },
    "фактическая_дата_выполнения": {
      "type": "date",
      "format": "YYYY-mm-dd"
    },
    "стоимость_заказа": {
      "type": "integer"
    },
    "id_брузады": {
      "type": "integer"
    },
  }
}
```

Анализатор

```
index_settings = {
  "settings": {
    "analysis": {
      "filter": {
        "my_stopwords": {
          "type": "stop",
          "stopwords": "_russian_"
        },
        "my_snowball": {
          "type": "snowball",
          "language": "Russian"
        }
      },
      "analyzer": {
        "my_custom_analyzer": {
          "type": "custom",
          "tokenizer": "standard",
          "filter": [
            "lowercase",
            "my_stopwords",
            "my_snowball"
          ]
        }
      }
    }
  }
}
```

Малпунг "Брузада"

```
mapping = {
  "properties": {
    "сведения_о_брузаде": {
      "type": "text",
      "analyzer": "my_custom_analyzer",
      "fielddata": True
    },
    "член_брузады": {
      "type": "text",
      "fielddata": True
    },
    "#analyzer": "my_custom_analyzer"
  },
  "отзыб_о_работе": {
    "type": "text",
    "analyzer": "my_custom_analyzer",
    "fielddata": True
  },
}
```

Разработка макета аналитической системы				Строительная компания						
Изм	Лист	№ докум.	Подпись	Дата	ElasticSearch Маллинг и анализатор			Лит	Масса	Масштаб
Разраб.		Горский Г.Е.			Лист 2			Листов		
Провер.		Григорьев Ю.А.								
					МГТУ им. Баумана ИУ6-23М					





					<i>Разработка макета аналитической системы строительная компания</i>					
					<i>Алгоритм для индексирования документов</i>	<i>Лит.</i>			<i>Масса</i>	<i>Масштаб</i>
<i>Изм.</i>	<i>Лист</i>	<i>№ докум.</i>	<i>Подп.</i>	<i>Дата</i>		<i>Лист 3</i>			<i>Листов</i>	
<i>Разраб.</i>		<i>Горский Г.Е.</i>				<i>МГТУ им. Баумана</i>				
<i>Проверил</i>		<i>Григорьев Ю.А.</i>				<i>ИУ6-23</i>				

Запрос: разбить заказы по дате заказа с периодом 1 год, для каждой группы определить суммарную стоимость заказов по каждой бригаде

Запрос:

```
GET orders/_search
{
  "size": 0,
  "aggs": {
    "orders_by_year": {
      "date_histogram": {
        "field": "дата_заказа",
        "calendar_interval": "1y"
      },
      "aggs": {
        "cost_by_team": {
          "terms": {
            "field": "id_бригады"
          },
          "aggs": {
            "total_cost": {
              "sum": {
                "field": "стоимость_заказа"
              }
            }
          }
        }
      }
    }
  }
}
```

Результат выполнения запроса:

```
{
  "took": 976,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 20,
      "relation": "eq"
    },
    "max_score": null,
    "hits": []
  },
  "aggregations": {
    "orders_by_year": {
      "buckets": [
        {
          "key_as_string": "2019-00-01",
          "key": 15463008000000,
          "doc_count": 5,
          "cost_by_team": {
            "doc_count_error_upper_bound": 0,
            "sum_other_doc_count": 0,
            "buckets": [
              {
                "key": 15,
                "doc_count": 2,
                "total_cost": {
                  "value": 9904.0
                }
              },
              {
                "key": 12,
                "doc_count": 1,
                "total_cost": {
                  "value": 7571.0
                }
              },
              {
                "key": 17,
                "doc_count": 1,
                "total_cost": {
                  "value": 8527.0
                }
              },
              {
                "key": 26,
                "doc_count": 1,
                "total_cost": {
                  "value": 6289.0
                }
              }
            ]
          }
        }
      ]
    }
  }
}
```

					Разработка макета аналитической системы строительная компания											
					ElasticSearch Первый запрос						Лит.		Масса	Масштаб		
Изм	Лист	№ докум.	Подп.	Дата												
Разраб.		Горский Г.Е.														
Проверил		Григорьев Ю.А.														
											Лист 4		Листов			
											МГТУ им. Баумана					
											ИУ6-23М					

Запрос: определить бригады хотя бы с одним отрицательным отзывом:

Запрос:

Результат выполнения запроса:

GET teams/\_search

{

"size": 0,

"aggs": {

"negative\_reviews": {

"filter": {

"terms": {

"отзыв\_о\_работе": ["вздвогнут",

"карма", "неправд"]

}

},

"aggs": {

"teams\_with\_negative": {

"terms": {

"field": "\_id",

"min\_doc\_count": 1,

"size": 60

}

}

}

}

}

}

{

"took": 118,

"timed\_out": false,

"\_shards": {

"total": 1,

"successful": 1,

"skipped": 0,

"failed": 0

},

"hits": {

"total": {

"value": 30,

"relation": "eq"

},

"max\_score": null,

"hits": []

},

"aggregations": {

"negative\_reviews": {

"doc\_count": 7,

"teams\_with\_negative": {

"doc\_count\_error\_upper\_bound": 0,

"sum\_other\_doc\_count": 0,

"buckets": [

{

"key": "11",

"doc\_count": 1

},

{

"key": "14",

"doc\_count": 1

},

{

"key": "16",

"doc\_count": 1

},

{

"key": "24",

"doc\_count": 1

},

{

"key": "28",

"doc\_count": 1

},

{

"key": "3",

"doc\_count": 1

},

{

"key": "6",

"doc\_count": 1

}

]

}

}

}

}

					Разработка макета аналитической системы строительная компания									
					ElasticSearch Второй запрос					Лит.		Масса	Масштаб	
Изм	Лист	№ докум.	Подп.	Дата										
Разраб.	Горский Г.Е.													
Проверил	Григорьев Ю.А.													
										Лист 5		Листов		
										МГТУ им. Баумана ИУ6-23М				








Запрос: найти компьютер, который принёс наибольший доход от продаж

Код запроса

```
MATCH (o:Order)-[p:Performed]->(t:Team)
WHERE p.actual_date < p.lead_date
RETURN o.name, p.lead_date, p.actual_date, t.id, t.team_info;
```

РЕЗУЛЬТАТ

\$ MATCH (o:Order)-[p:Performed]->(t:Team) WHERE p.actual_date < p.lead_date RETURN o.name, p.lead_date, p.actual_d...						
 Table	o.name	p.lead_date	p.actual_date	t.id	t.team_info	
 Text	52675	"2022-12-15"	"2022-01-10"	"20"	["неудобн", "исполня", "песенк", "правильн", "сомнительн", "точн", "миф", "мелоч", "тысяч", "упор", "очередн", "недостаток", "летет", "еврейск", "песн", "девк", "заплака", "устройств"]	
	22603	"2022-05-12"	"2021-11-19"	"13"	["монет", "коробк", "зарплат", "стен", "привлека", "одиннадц", "карма", "пищ", "сход"]	
 Code	3624	"2023-02-04"	"2021-04-28"	"15"	["устройств", "миг", "нож", "господ", "возбужден", "июн", "налев", "процесс", "армейск", "бригад"]	
	51612	"2022-04-14"	"2021-08-26"	"17"	["пол", "народ", "салон", "наста", "наступа", "господ", "житель", "палец", "выражен", "стен", "ответ", "песенк", "выгна", "растеря", "набор"]	
Started streaming 4 records after 451 ms and completed after 562 ms.						

Разработка макета аналитической системы									
Строительная компания									
Запрос на языке Cypher				Лист		Масса		Масштаб	
				Лист 7		Листов			
МГТУ им. Баумана									
ИУ6-23М									





					Разработка макета аналитической системы строительная компания				
					Алгоритм создания csv-файлов				
Изм.	Лист	№ докум.	Подп.	Дата					
Разраб.	Горский Г.Е.								
Проверил	Григорьев Ю.А.								
					Лит.      Масса      Масштаб				
					Лист 8      Листов				
					МГТУ им. Баумана				
					ИУ6-23				



# Запрос: найти суммарную стоимость заказов по каждой заказчику

## КОД ЗАПРОСА

```
from pyspark.sql import SparkSession

ss = SparkSession.builder.appName('read_csv').getOrCreate()
ss.read.options(header='true').csv('hdfs://localhost:9000/1/orders.csv').createOrReplaceTempView('orders')
ss.read.options(header='true').csv('hdfs://localhost:9000/1/customers.csv').createOrReplaceTempView('customers')

res_df = ss.sql("""
select c.customer_id, c.customer_info, SUM(o.cost)
from orders o
join customers c on c.customer_id = o.customer_id
GROUP BY 1,2
ORDER BY 3 desc
""")

res_df.show()
input("Ctrl C")
```

## РЕЗУЛЬТАТ

customer_id	customer_info	sum(CAST(cost AS DOUBLE))
27005	медведев агар вик...	24237.0
82279	алин тимофеевн ко...	15388.0
64694	артёмьев ксен рус...	9349.0
38137	папф бориславович...	9175.0
5374	козлов таис вадимовн	8586.0
629	буров нинел рудол...	8527.0
10107	игор венедиктович...	7571.0
98001	лор николаевн мак...	6289.0
42667	анастас станицлав...	6241.0
77033	ершов варвар вени...	6014.0
74103	аксенов маргарит ...	5321.0
53991	станимир викентье...	5041.0
79925	герма федотович тит	3612.0
25950	сазон офиног елиз...	3555.0
31832	сафонов антонин в...	2941.0
16179	вероник владислав...	1318.0
98465	галин михайловн я...	37.0

Разработка макета аналитической системы Строительная компания				Лист		Масса	Масштаб
				Лист 9		Листов	
Spark. Запрос							
МГТУ им. Баумана ИУБ-23М							



Spark. Монитор

Выполнение SQL-запроса



Summary Metrics for 2 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	0,6 s	0,6 s	0,6 s	0,6 s	0,6 s
GC Time	0 ms	0 ms	0 ms	0 ms	0 ms
Input Size / Records	541,0 B / 10	541,0 B / 10	556,0 B / 10	556,0 B / 10	556,0 B / 10
Shuffle Write Size / Records	902,0 B / 7	902,0 B / 7	1208,0 B / 10	1208,0 B / 10	1208,0 B / 10

Aggregated Metrics by Executor

Executor ID	Address	Task Time	Total Tasks	Failed Tasks	Killed Tasks	Succeeded Tasks	Input Size / Records	Shuffle Write Size / Records	Blacklisted
driver	localhost:39395	2 s	2	0	0	2	1097,0 B / 20	2,1 KB / 17	false

Tasks (2)

Index	ID	Attempt	Status	Locality Level	Executor ID	Host	Launch Time	Duration	GC Time	Input Size / Records	Write Time	Shuffle Write Size / Records	Errors
0	4	0	SUCCESS	ANY	driver	localhost	2023/05/18 01:02:41	0,6 s		556,0 B / 10	0,1 s	902,0 B / 7	
1	5	0	SUCCESS	ANY	driver	localhost	2023/05/18 01:02:41	0,6 s		541,0 B / 10	0,1 s	1208,0 B / 10	

Выполненные SQL-запросы

SQL

Completed Queries: 5

Completed Queries (5)

ID	Description	Submitted	Duration	Job IDs
4	showString at NativeMethodAccessorImpl.java:0	2023/05/18 01:02:33	18 s	[1]
3	createOrReplaceTempView at NativeMethodAccessorImpl.java:0	2023/05/18 01:02:32	0 ms	
2	csv at NativeMethodAccessorImpl.java:0	2023/05/18 01:02:31	0,3 s	[1]
1	createOrReplaceTempView at NativeMethodAccessorImpl.java:0	2023/05/18 01:02:31	7 ms	
0	csv at NativeMethodAccessorImpl.java:0	2023/05/18 01:02:23	6 s	[0]

"Event Timeline" скрипта с SQL-запросом

Event Timeline

Enable zooming

Executors

Added

Removed

Jobs

Succeeded

Failed

Running



Completed Jobs (4)

Job ID	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
3	showString at NativeMethodAccessorImpl.java:0	2023/05/18 01:02:41	10 s	2/2	200/200
2	run at ThreadPoolsExecutor.java:1149	2023/05/18 01:02:38	2 s	1/1	3/3
1	csv at NativeMethodAccessorImpl.java:0	2023/05/18 01:02:32	0,1 s	1/1	1/1
0	csv at NativeMethodAccessorImpl.java:0	2023/05/18 01:02:27	2 s	1/1	1/1

Разработка макета аналитической системы строительная компания

Spark. Монитор

Лит.

Масса

Масштаб

Изм. Лист № докум. Подп. Дата

Разраб. Горский Г.Е.

Проверил Григорьев Ю.А.

Лист 10

Листов

МГТУ им. Баумана

ИУ6-23