

Тема курсового проекта:

Разработка макета аналитической системы

Вариант №1: Ремонт бытовой техники**Elasticsearch**

1. Необходимо сформировать два типа JSON-документов со следующими полями:

Заказ:

```
{index, doc_type, id, body:
{id_заказа, дата_заказа, id_заказчика, сведения_o_заказчике*, данные_o_заказе*,
срок_выполнения_заказа, фактическая_дата_выполнения, [запчасть*], [ремонт*],
стоимость, id_мастера}}
```

Мастер:

```
{index, doc_type, id, body: {сведения_o_мастере*, [отзыв_o_работе*]}}
```

Примечание. Квадратные скобки [] обозначает тег (может быть несколько значений)

2. Требование к анализатору:

Поля, отмеченные *, разделить на слова, убрать пунктуацию с помощью токенизатора standart (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 месяц, для каждой «корзины» определить суммарное число заказов по каждой запчасти,
- вывести мастеров, в сведениях которых указан стаж работы.

Neo4j

1. По данным из Elasticsearch заполнить графовую базу данных:

Заказ(id_заказа, дата_заказа, сведения_o_заказчике, стоимость_заказа) -
Выполнил(срок_выполнения_заказа, фактическая_дата_выполнения) - Мастер(id_мастера, сведения_o_мастере).

Примечание. В скобках приведены свойства узлов и отношения (связи), глагол – это отношение.

2. Разработать и реализовать запрос:

Найти мастера, который выполнил максимальное количество заказов.

Spark

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

2. Написать запрос select: найти заказы и мастеров, которые не выполнили заказы в срок.

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

Согласовано				- вывести мастеров, в сведениях которых указан стаж работы.									
	Neo4j												
	1. По данным из Elasticsearch заполнить графовую базу данных: Заказ(id_заказа, дата_заказа, сведения_o_заказчике, стоимость_заказа) - Выполнил(срок_выполнения_заказа, фактическая_дата_выполнения) - Мастер(id_мастера, сведения_o_мастере). <u>Примечание.</u> В скобках приведены свойства узлов и отношения (связи), глагол – это отношение.												
	2. Разработать и реализовать запрос: Найти мастера, который выполнил максимальное количество заказов.												
Взам. инб №	Spark												
	1. По данным из Elasticsearch сформировать csv-файлы (с внутренней схемой) таблиц «Заказчик», «Заказ», «Мастер» и сохранить их в файловой системе HDFS.												
Подп. и дата	2. Написать запрос select: найти заказы и мастеров, которые не выполнили заказы в срок.												
	3. Реализовать этот запрос в Spark. Построить временную диаграмму его выполнения по результатам работы монитора.												
Инб № подл.							Разработка макета аналитической системы (Вариант №1)						
	Изм.	Колуч	Лист	№ док	Подп.	Дата	Название темы курсового проекта, задание и описание варианта			Стадия	Лист	Листов	
	Разраб.	Астахов С.В.									1	11	
	Рукоб.	Григорьев Ю.А.											
							МГТУ им. Н.Э. Баумана Группа ИУ6-22М						
Н. Контр.													

Индексация документов Elasticsearch

Маппинг для типа "Мастер":

```
mappings_master = {
  "mappings" : {
    "properties" : {
      "master_desc" : {
        "type" : "text",
        "analyzer": "custom_analyzer"
      },
      "master_feedbacks" : {
        "type" : "text",
        "fields" : {
          "keyword" : {
            "type" : "keyword"
          }
        },
        "analyzer": "custom_analyzer"
      },
      "master_id" : {
        "type" : "long"
      }
    }
  }
}
```

Маппинг для типа "Заказ":

```
mappings_order = {
  "mappings" : {
    "properties" : {
      "order_customer_desc" : {
        "type" : "text",
        "analyzer": "custom_analyzer"
      },
      "order_customer_id" : {
        "type" : "long"
      },
      "order_date" : {
        "type" : "date"
      },
      "order_details_desc" : {
        "type" : "text",
        "analyzer": "custom_analyzer"
      },
      "order_due_date" : {
        "type" : "date"
      },
      "order_fact_completion_date" : {
        "type" : "date"
      },
      "order_id" : {
        "type" : "long"
      },
      "order_master_id" : {
        "type" : "long"
      },
      "order_parts" : {
        "type" : "text",
        "fielddata" : True,
        "fields" : {
          "keyword" : {"type" : "keyword"}
        },
        "analyzer": "custom_analyzer"
      },
      "order_price" : {
        "type" : "float"
      },
      "repair_types" : {
        "type" : "text",
        "fields" : {
          "keyword" : {"type" : "keyword"}
        },
        "analyzer": "custom_analyzer"
      }
    }
  }
}
```

Конфигурация анализатора:

```
analyzer_settings = {
  "settings": {
    "analysis": {
      "filter": {
        "ru_stop": {
          "type": "stop",
          "stopwords": "_russian_"
        },
        "snow_ru_stemmer": {
          "type": "snowball",
          "language": "russian"
        }
      },
      "analyzer": {
        "custom_analyzer": {
          "type": "custom",
          "tokenizer": "standard",
          "filter": [
            "lowercase",
            "ru_stop",
            "snow_ru_stemmer"
          ]
        }
      }
    }
  }
}
```

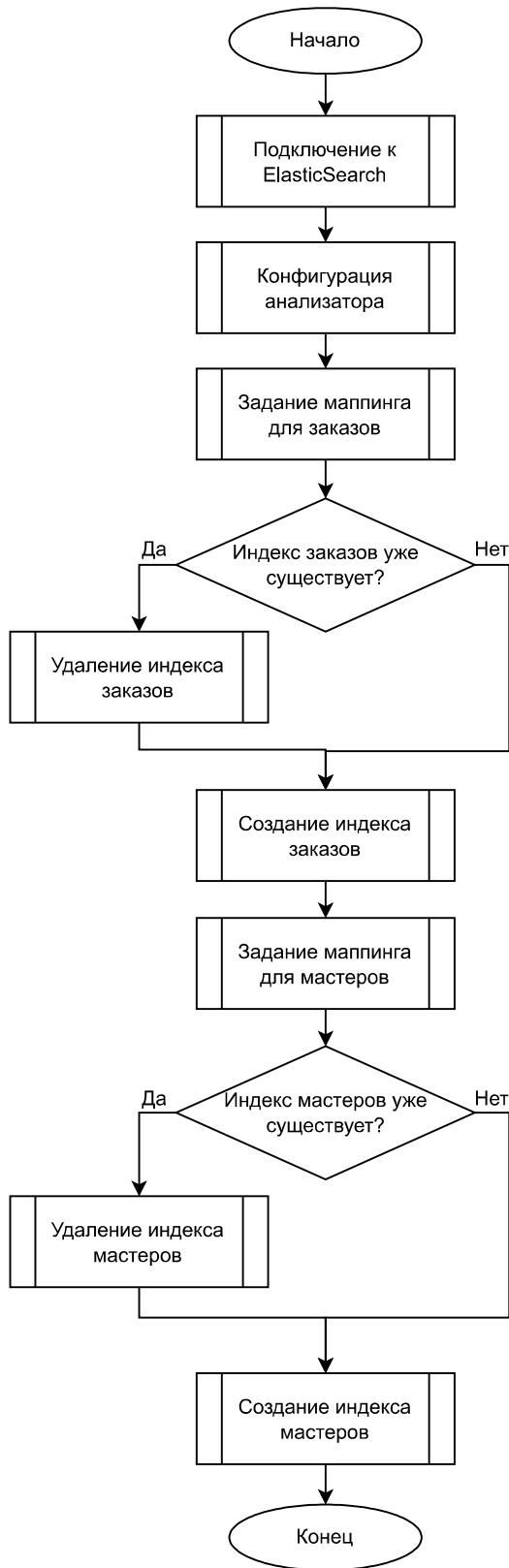
Согласовано			
Взам. инб. №			
Подп. и дата			
Инб. № подл.	Разраб.	Астахов С.В.	
	Руков.	Григорьев Ю.А.	
	Н. Контр.		

						Разработка макета аналитической системы (Вариант №1)			
Изм.	Колуч	Лист	№ док	Подп.	Дата				
Разраб.		Астахов С.В.				Стадия		Лист	Листов
Руков.		Григорьев Ю.А.						2	11
						МГТУ им. Н.Э. Баумана Группа ИУ6-22М			
Н. Контр.									
Elasticsearch									
Описание анализатора и маппинга									

Elasticsearch
Описание анализатора и маппинга

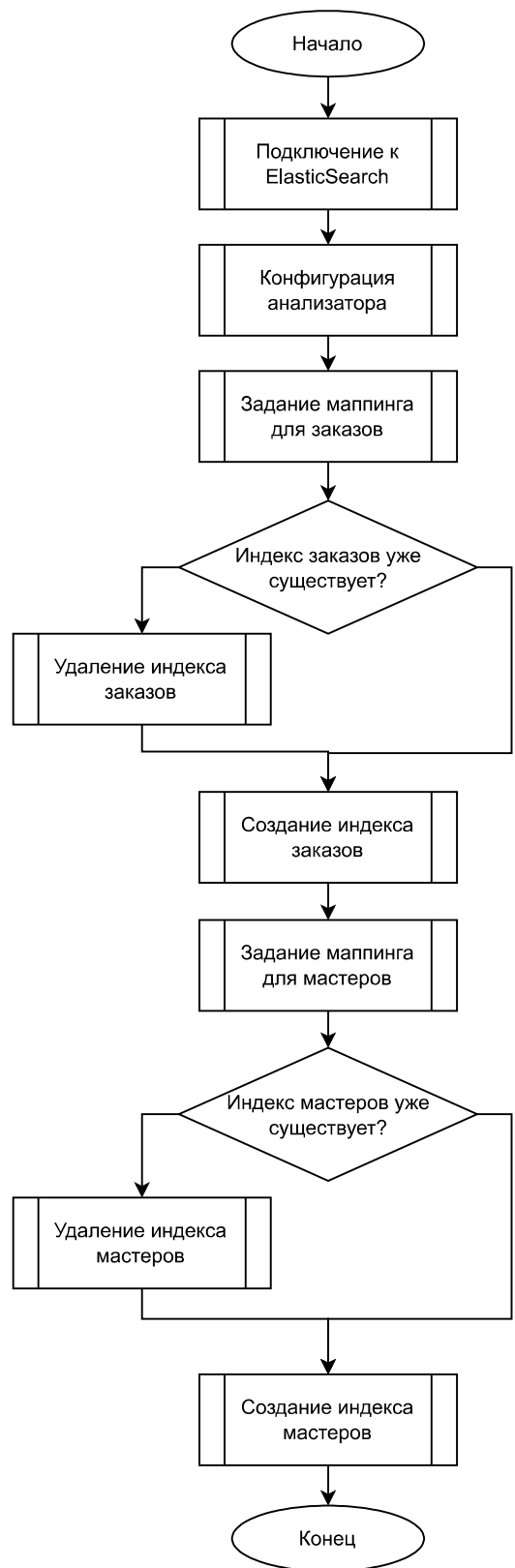
МГТУ им. Н.Э. Баумана
Группа ИУ6-22М

Схема алгоритма программы индексации



Согласовано			
Взам. инб. №			
Подп. и дата			
Инб. № подл.			

						Разработка макета аналитической системы (Вариант №1)			
Изм.	Колуч	Лист	№ док	Подп.	Дата				
Разраб.	Астахов С.В.					Elasticsearch Схема алгоритма программы индексации документов	Стадия	Лист	Листов
Руков.	Григорьев Ю.А.						3	11	
							МГТУ им. Н.Э. Баумана Группа ИУ6-22М		
Н. Контр.									



Текст и результаты выполнения запросов

Текст запроса №1:

```
GET order/_search
{
  "aggs": {
    "over_months": {
      "date_histogram": {
        "field": "order_date",
        "calendar_interval": "month",
        "format": "yyyy-MM-dd"
      },
      "aggs": {
        "over_parts": {
          "terms": {
            "field": "order_parts.keyword"
          }
        }
      }
    }
  }
}
```

Результат выполнения запроса №1:

```
[{'key_as_string': '2023-12-01',
'key': 1701388800000,
'doc_count': 13,
'over_parts': {'doc_count_error_upper_bound': 0,
'sum_other_doc_count': 0,
'buckets': [{'key': 'usb-разъем', 'doc_count': 5},
{'key': 'дисплей', 'doc_count': 4},
{'key': 'корпус', 'doc_count': 4},
{'key': 'аккумулятор', 'doc_count': 3}]}],
{'key_as_string': '2024-01-01',
'key': 1704067200000,
'doc_count': 48,
'over_parts': {'doc_count_error_upper_bound': 0,
'sum_other_doc_count': 0,
'buckets': [{'key': 'usb-разъем', 'doc_count': 23},
{'key': 'дисплей', 'doc_count': 22},
{'key': 'корпус', 'doc_count': 18},
{'key': 'аккумулятор', 'doc_count': 16}]}],
{'key_as_string': '2024-02-01',
'key': 1706745600000,
'doc_count': 37,
'over_parts': {'doc_count_error_upper_bound': 0,
'sum_other_doc_count': 0,
'buckets': [{'key': 'дисплей', 'doc_count': 19},
{'key': 'usb-разъем', 'doc_count': 14},
{'key': 'аккумулятор', 'doc_count': 14},
{'key': 'корпус', 'doc_count': 14}]}],
{'key_as_string': '2024-03-01',
'key': 1709251200000,
'doc_count': 2,
'over_parts': {'doc_count_error_upper_bound': 0,
'sum_other_doc_count': 0,
'buckets': [{'key': 'usb-разъем', 'doc_count': 1},
{'key': 'аккумулятор', 'doc_count': 1},
{'key': 'корпус', 'doc_count': 1}]}]}
```

Текст запроса №2:

```
GET master/_search
{
  "query": {
    "match": {
      "master_desc": "стажа"
    }
  }
}
```

Согласовано				<div>Текст запроса №2:</div> <div>GET master/_search { "query": { "match": { "master_desc": "стажа" } } }</div>		<div>doc_count': 57, 'over_parts': {'doc_count_error_upper_boun d': 0, 'sum_other_doc_count': 0, 'buckets': [{'key': 'дисплей', 'doc_coun t': 19}, {'key': 'usb-разъем', 'doc_count': 14}, {'key': 'аккумулятор', 'doc_count': 14}, {'key': 'корпус', 'doc_count': 14}]}, {'key_as_string': '2024-03-01', 'key': 1709251200000, 'doc_count': 2, 'over_parts': {'doc_count_error_upper_boun d': 0, 'sum_other_doc_count': 0, 'buckets': [{'key': 'usb-разъем', 'doc_co unt': 1}, {'key': 'аккумулятор', 'doc_count': 1}, {'key': 'корпус', 'doc_count': 1}]}}}</div>									
Взам. инв. №															
Подп. и дата															
Инв. № подл.	Изм.	Колуч	Лист	№ док	Подп.	Дата	Разработка макета аналитической системы (Вариант №1)			Elasticsearch Текст и результаты выполнения запросов			Стадия	Лист	Листов
	Разраб.	Астахов С.В.												4	11
	Руков.	Григорьев Ю.А.													
	Н. Контр.														
МГТУ им. Н.Э. Баумана Группа ИУ6-22М															

Текст и результаты выполнения запросов

Результат выполнения запроса №2:

```
{'total': {'value': 8, 'relation': 'eq'},
 'max_score': 0.8435577,
 'hits': [{'_index': 'master',
            '_type': '_doc',
            '_id': '22653',
            '_score': 0.8435577,
            '_source': {'master_id': 22653,
                        'master_desc': 'Акулина Рудольфовна Никитина, Стаж Работы: 15 л./г..',
                        'master_feedbacks': ['ворчливый, аккуратный.',
                                             'ворчливый, медлительный.']}},
          {'_index': 'master',
            '_type': '_doc',
            '_id': '300714',
            '_score': 0.8435577,
            '_source': {'master_id': 300714,
                        'master_desc': 'Вероника Петровна Силина, Стаж Работы: 12 л./г..',
                        'master_feedbacks': ['ворчливый, медлительный.',
                                             'ворчливый, аккуратный.']}},
          {'_index': 'master',
            '_type': '_doc',
            '_id': '405064',
            '_score': 0.8435577,
            '_source': {'master_id': 405064,
                        'master_desc': 'Хохлов Олег Харлампьевич, Стаж Работы: 3 л./г..',
                        'master_feedbacks': ['ворчливый, медлительный.',
                                             'ворчливый, аккуратный.']}},
          {'_index': 'master',
            '_type': '_doc',
            '_id': '257458',
            '_score': 0.8435577,
            '_source': {'master_id': 257458,
                        'master_desc': 'Шестакова Наина Владиславовна, Стаж Работы: 9 л./г..',
                        'master_feedbacks': ['аккуратный, ворчливый.', 'ворчливый, аккуратный.']}},
          {'_index': 'master',
            '_type': '_doc',
            '_id': '814777',
            '_score': 0.8435577,
            '_source': {'master_id': 814777,
                        'master_desc': 'Шарова Жанна Аркадьевна, Стаж Работы: 1 л./г..',
                        'master_feedbacks': ['ворчливый, медлительный.',
                                             'медлительный, аккуратный.',
                                             'ворчливый, медлительный.']}},
          {'_index': 'master',
            '_type': '_doc',
            '_id': '847121',
            '_score': 0.8435577,
            '_source': {'master_id': 847121,
                        'master_desc': 'Жуков Владимир Всеволодович, Стаж Работы: 12 л./г..',
                        'master_feedbacks': ['медлительный, ворчливый.',
                                             'аккуратный, медлительный.']}}, ...
        ]}
```

Согласовано		

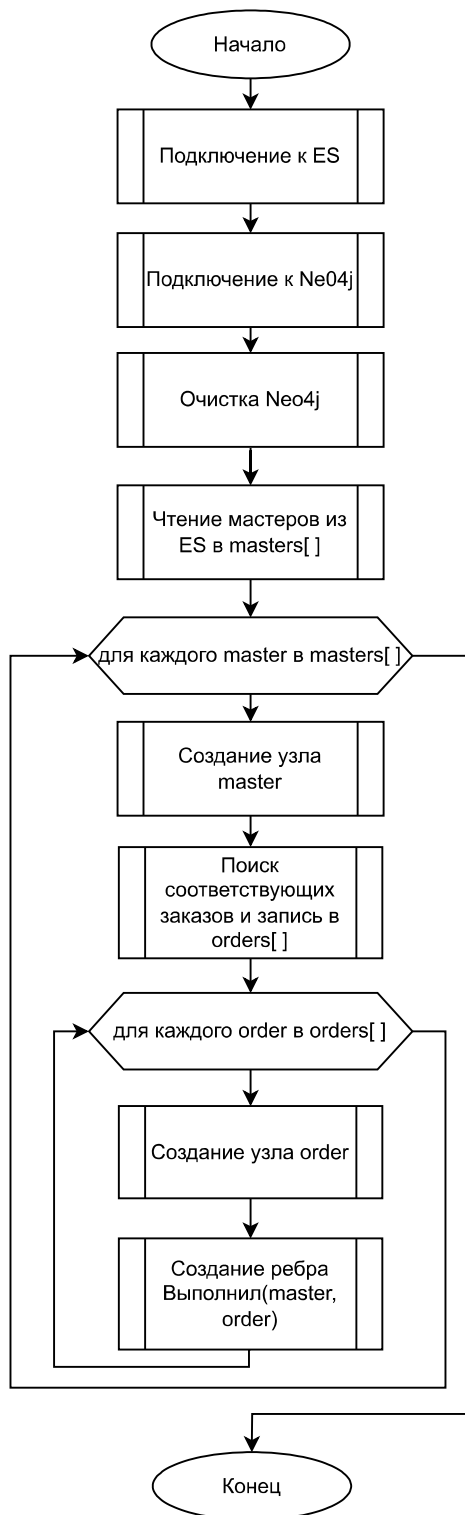
Взам. инб №

Подп. и дата

Инб № подл.

						Разработка макета аналитической системы (Вариант №1)			
Изм.	Колуч	Лист	№ док	Подп.	Дата				
Разраб.		Астахов С.В.				Elasticsearch Текст и результаты выполнения запросов	Стадия	Лист	Листов
Руков.		Григорьев Ю.А.						5	11
							МГТУ им. Н.Э. Баумана Группа ИУ6-22М		
Н. Контр.									

Схема алгоритма создания и заполнения графовой базы данных



Согласовано			

Взам. инб. №

Подп. и дата

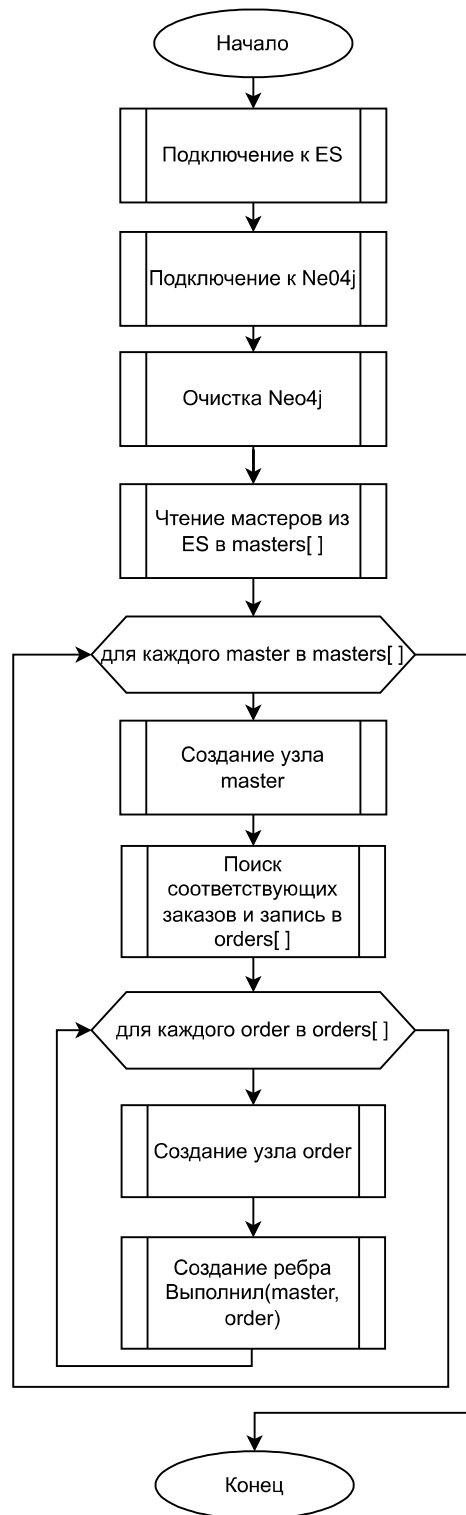
Инб. № подл.

Изм.	Колуч	Лист	№ док	Подп.	Дата
Разраб.	Астахов С.В.				
Руков.	Григорьев Ю.А.				
Н. Контр.					

Разработка макета аналитической системы (Вариант №1)

Neo4j
Схема алгоритма создания и заполнения
графовой базы данных

Стадия	Лист	Листов
	6	11
МГТУ им. Н.Э. Баумана Группа ИУ6-22М		



Текст запроса на языке Cypher и результаты выполнения запроса

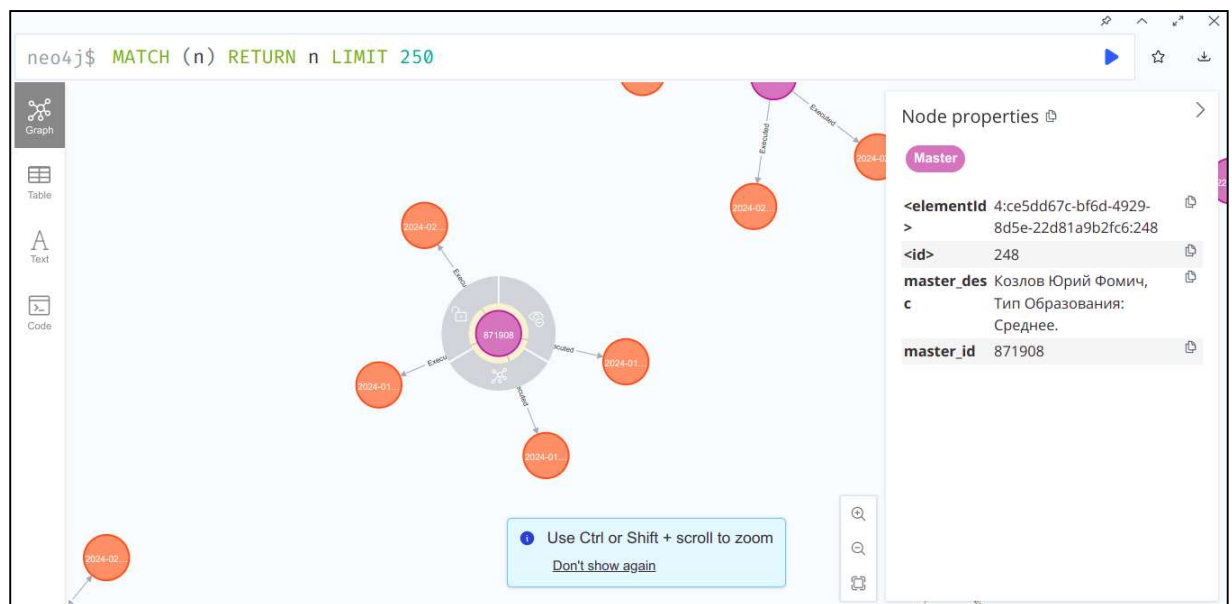
Текст запроса на языке Cypher:

```
MATCH (MAS:Master)-[r:Executed]->(ORD:Order) WITH MAS, count(r) AS num
RETURN MAS.master_desc as master_desc, num ORDER BY num DESC LIMIT 1;
```

Результат выполнения запроса:

'Воробьев Ювеналий Измаилович, Тип Образования: Среднее.' 10

Фрагмент визуализации исходного графа:



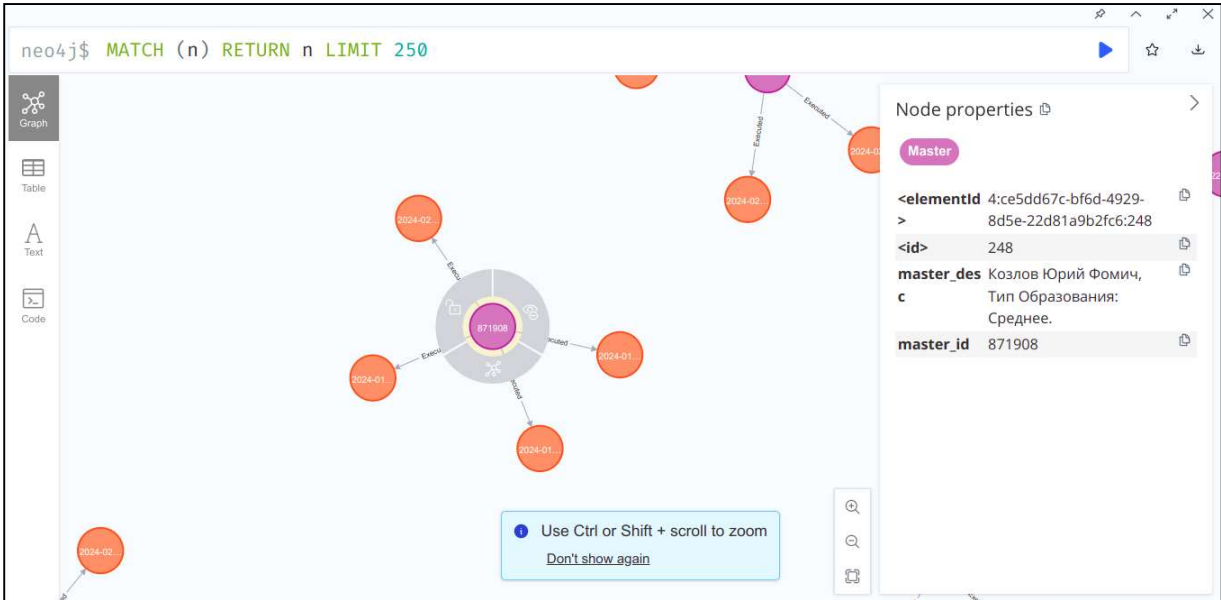
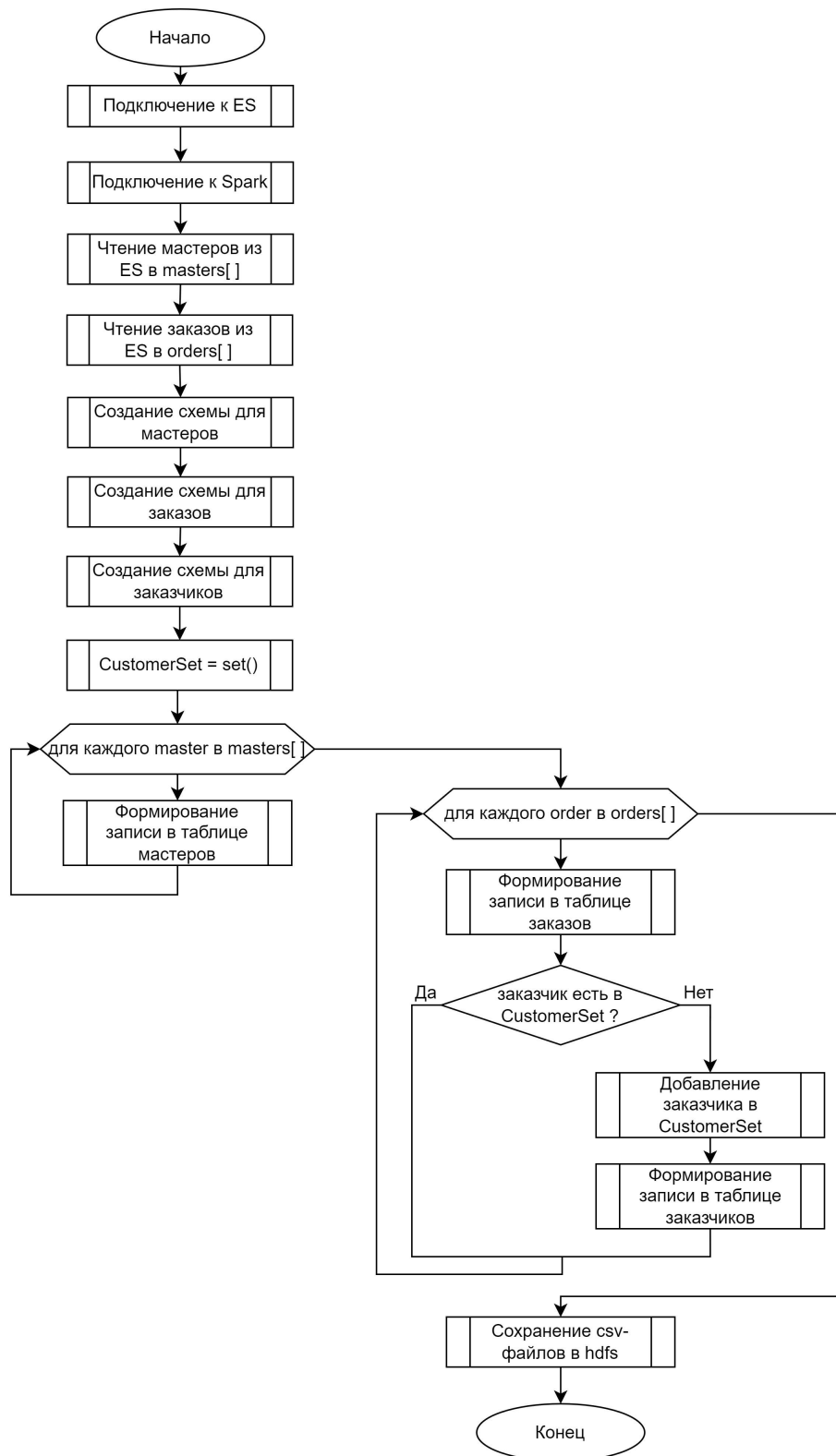
Согласовано							<div><div>neo4j\$ MATCH (n) RETURN n LIMIT 250</div><div><div><div>Graph</div><div>Table</div><div>Text</div><div>Code</div></div><div></div></div></div>					
	Взам. инб. №							Разработка макета аналитической системы (Вариант №1)				
Подп. и дата							Изм.КолучЛист№ докПодп.Дата					
Инб. № подл.	Разраб.	Астахов С.В.					Neo4j Текст запроса на языке Cypher Результаты выполнения запроса			Стадия	Лист	Листов
	Руков.	Григорьев Ю.А.									7	11
										МГТУ им. Н.Э. Баумана Группа ИУ6-22М		
	Н. Контр.											

Схема алгоритма создания CSV-файлов



Согласовано			

Взам. инб. №

Подп. и дата

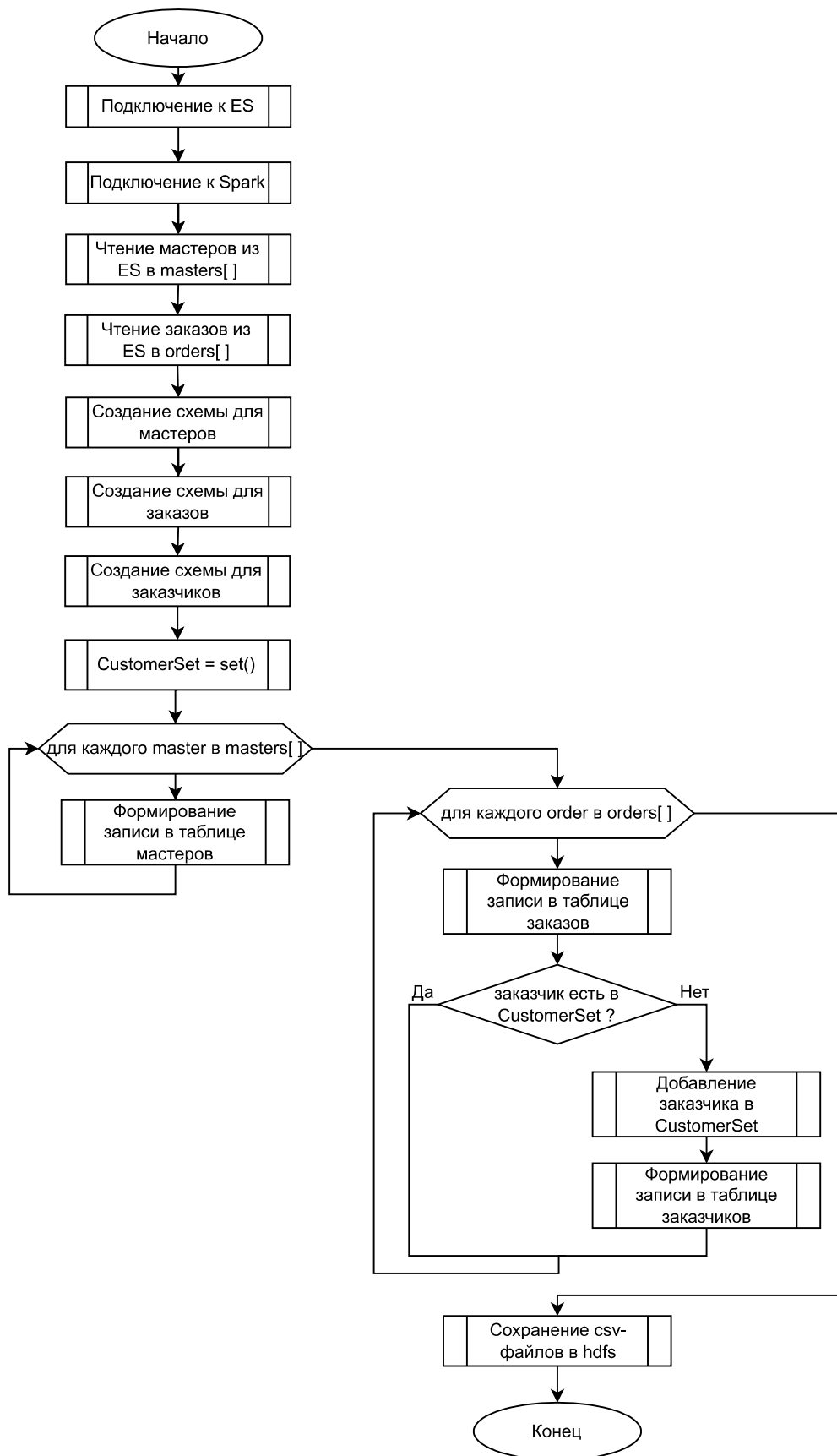
Инб. № подл.

Изм.	Колуч	Лист	№ док	Подп.	Дата
Разраб.		Астахов С.В.			
Руков.		Григорьев Ю.А.			
Н. Контр.					

Разработка макета аналитической системы (Вариант №1)

Spark
Схема алгоритма создания CSV-файлов

Стадия	Лист	Листов
	8	11
МГТУ им. Н.Э. Баумана Группа ИУ6-22М		



Текст и результат запроса в Spark

Текст запроса:

```
select m.master_id, m.master_desc, o.order_id, o.order_date, o.order_due_date,
o.order_fact_completion_date, c.order_customer_id, c.order_customer_desc

from master m JOIN order o ON o.order_master_id = m.master_id LEFT JOIN customer c ON
o.order_customer_id = c.order_customer_id

where o.order_fact_completion_date > o.order_due_date
```

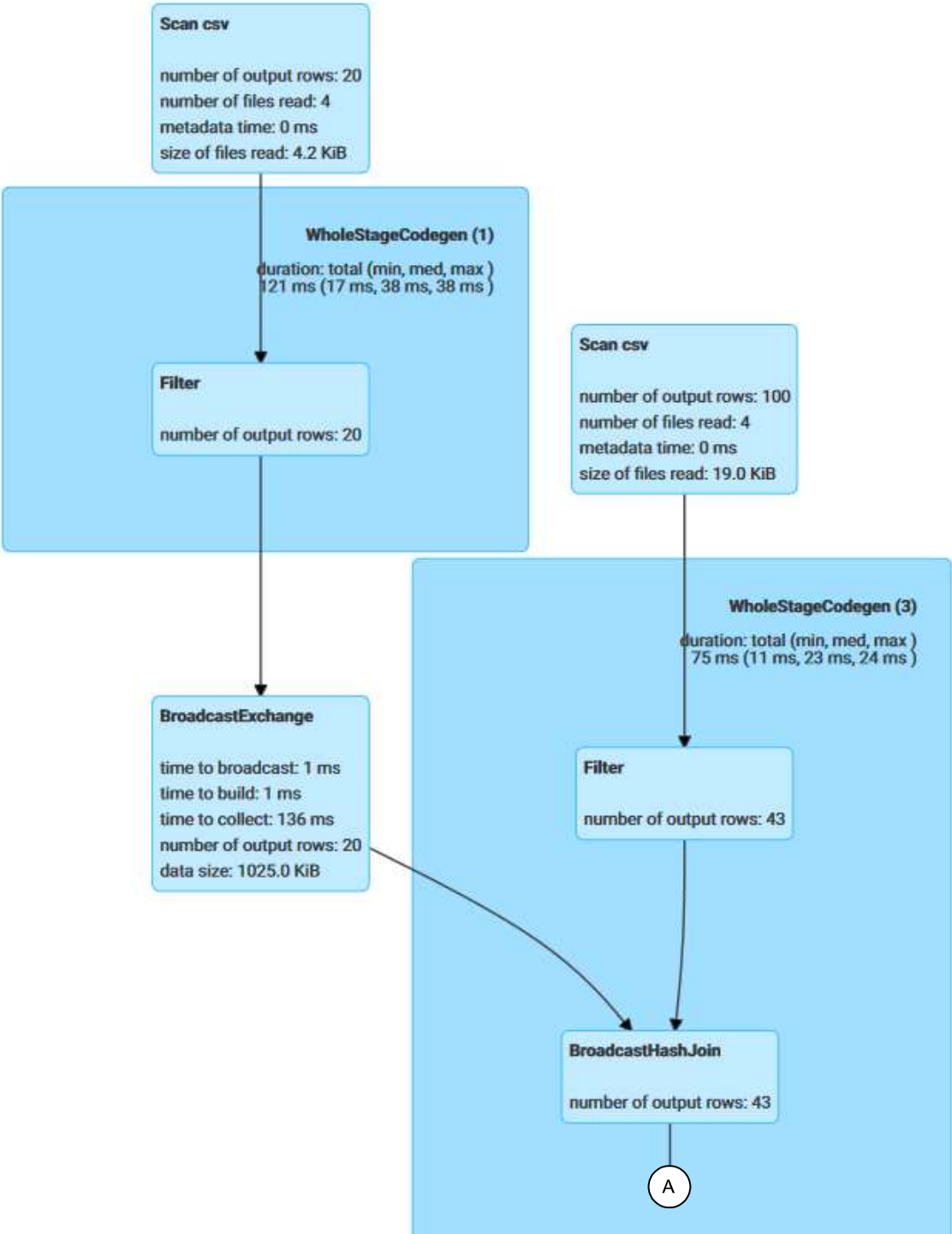
Текст запроса №2:

master_id	master_desc	order_id	order_date	due_date	fact_compl_date	customer_id	customer_desc
257458	Шестакова На...	131919	2024-01-06	2024-04-10	2024-04-27	227431	Имя: Маргари...
814777	Шарова Жанна...	141550	2024-01-10	2024-04-03	2024-04-27	683623	Имя: Лора Бо...
871908	Козлов Юрий ...	993858	2024-01-16	2024-04-08	2024-04-27	433504	Имя: Лукьян ...
300714	Вероника Пет...	683490	2024-01-22	2024-04-15	2024-04-20	269146	Имя: Алина О...
736774	Владимирова ...	243483	2024-02-01	2024-04-05	2024-04-27	408481	Имя: Филатов...
785701	Панфилова пе...	321753	2024-01-20	2024-04-05	2024-04-14	310962	Имя: Комаров...
563268	Милица Русла...	598097	2024-01-04	2024-04-07	2024-04-21	805728	Имя: Евфроси...
257458	Шестакова На...	794545	2024-01-28	2024-04-11	2024-04-28	730477	Имя: Лукин А...
464209	Ия Тарасовна...	543677	2024-01-31	2024-04-01	2024-04-26	987015	Имя: Роман Ф...
405064	Хохлов Олег ...	687025	2024-02-28	2024-04-11	2024-04-19	853109	Имя: Федосий...
847121	Жуков Ладими...	786535	2024-01-14	2024-04-01	2024-04-16	977695	Имя: Исай Фр...
164475	Юлий Дмитрие...	767369	2024-02-15	2024-04-12	2024-04-15	204174	Имя: Кудряшо...
563268	Милица Русла...	948558	2024-02-01	2024-04-23	2024-04-30	492289	Имя: Наталья...
22653	Акулина Рудо...	804565	2024-01-30	2024-04-06	2024-04-26	468141	Имя: Ангелин...
464209	Ия Тарасовна...	75572	2024-01-13	2024-04-08	2024-04-13	691138	Имя: Прохоро...
397809	Тарасова Фёк...	596544	2024-02-12	2024-04-04	2024-04-16	822221	Имя: Кудряшо...
282806	Маслов Емель...	161763	2024-02-13	2024-04-10	2024-04-29	608202	Имя: Гусев А...
163543	Максимильтян ...	271105	2024-01-09	2024-04-15	2024-04-19	569974	Имя: Серафим...
405064	Хохлов Олег ...	426371	2023-12-24	2024-04-10	2024-04-13	422303	Имя: Матвеев...
736774	Владимирова ...	371950	2024-01-24	2024-04-12	2024-04-14	507529	Имя: Глафира...
464209	Ия Тарасовна...	647404	2024-02-06	2024-04-03	2024-04-23	975300	Имя: Г-н Еме...
620880	Комиссаров К...	337041	2024-02-18	2024-04-04	2024-04-27	464422	Имя: Гордеев...
22653	Акулина Рудо...	366833	2024-02-20	2024-04-20	2024-04-23	830837	Имя: Горбаче...
394321	Воробьев Юве...	493503	2024-02-16	2024-04-05	2024-04-13	233350	Имя: Кузьмин...
257458	Шестакова На...	991702	2024-01-13	2024-04-19	2024-04-30	822548	Имя: Зинаида...
563268	Милица Русла...	665466	2024-01-18	2024-04-20	2024-04-21	363183	Имя: Сидоров...
164475	Юлий Дмитрие...	152939	2024-01-16	2024-04-07	2024-04-27	436264	Имя: тов. ме...
785701	Панфилова пе...	38299	2023-12-27	2024-04-16	2024-04-28	778168	Имя: Арефий ...
220028	Зиновий дими...	37803	2024-02-19	2024-04-17	2024-04-25	115391	Имя: Муравье...
620880	Комиссаров К...	950739	2024-02-10	2024-04-08	2024-04-30	441557	Имя: Никифор...
229678	Евпраксия Ал...	344458	2023-12-31	2024-04-23	2024-04-25	992807	Имя: Крюкова...
397809	Тарасова Фёк...	181936	2024-01-16	2024-04-08	2024-04-10	498893	Имя: Симонов...
871908	Козлов Юрий ...	8396	2024-01-19	2024-04-22	2024-04-28	306651	Имя: Зинаида...

Согласовано			
Взам. инб. №			
Подп. и дата			
Инб. № подл.			

						Разработка макета аналитической системы (Вариант №1)			
Изм.	Колуч	Лист	№ док	Подп.	Дата				
Разраб.		Астахов С.В.				Spark Текст и результат запроса	Стадия	Лист	Листов
Руков.		Григорьев Ю.А.						9	11
							МГТУ им. Н.Э. Баумана Группа ИУ6-22М		
Н. Контр.									

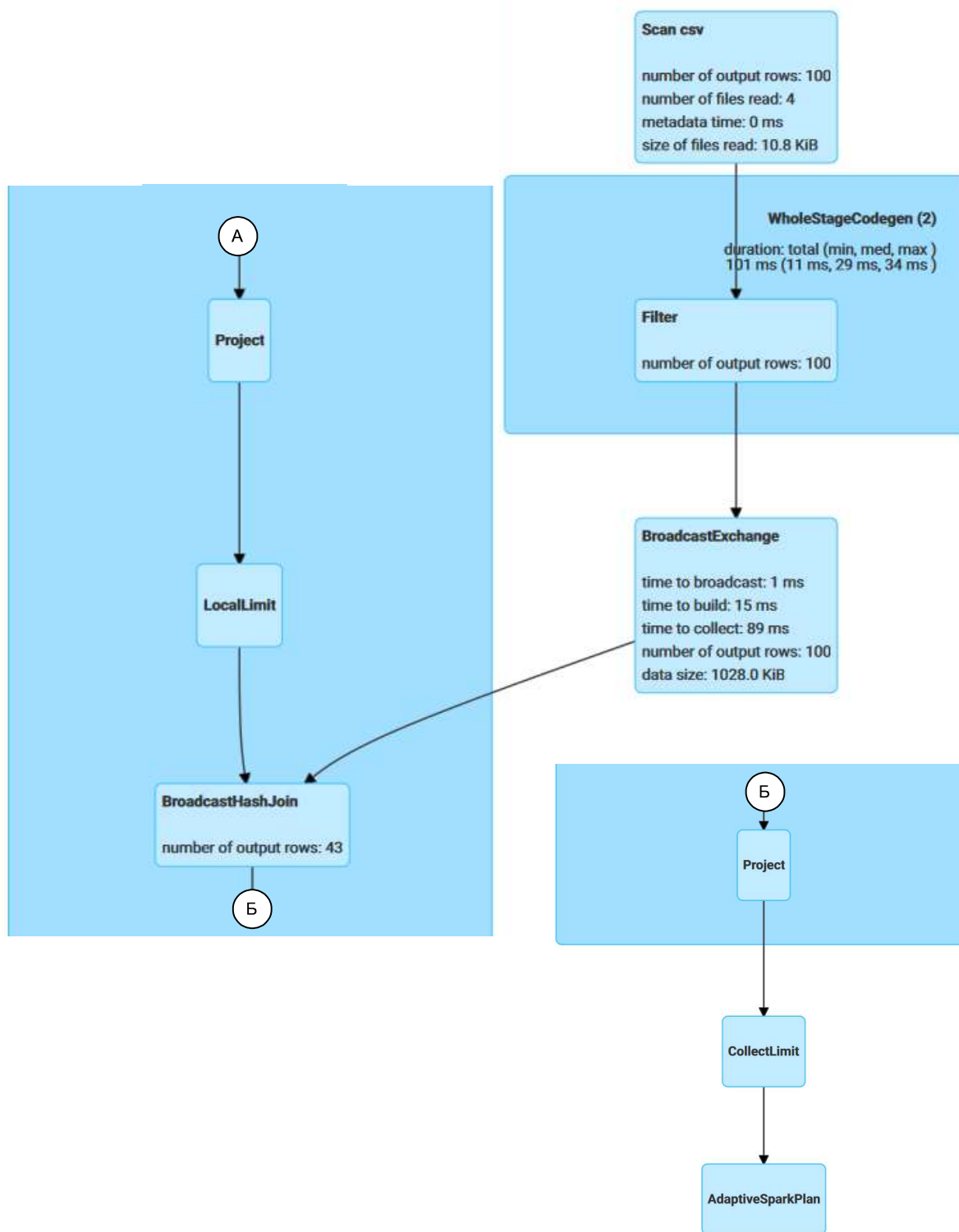
DAG выполнения запроса



Согласовано					
Взам. инб. №					
Подп. и дата					
Инб. № подл.					

						Разработка макета аналитической системы (Вариант №1)			
Изм.	Колуч	Лист	№ док	Подп.	Дата				
Разраб.	Астахов С.В.					Spark DAG выполнения запроса	Стадия	Лист	Листов
Руков.	Григорьев Ю.А.							10	11
							МГТУ им. Н.Э. Баумана Группа ИУ6-22М		
Н. Контр.									

DAG выполнения запроса (ч. 2)



Согласовано			

ВЗДМ. ЦНБ №

Հոփոն և ձառո

Инв № подл.

						Разработка макета аналитической системы (Вариант №1)			
Изм.	Колуч	Лист	№ док	Подп.	Дата	Spark DAG выполнения запроса (ч. 2)	Стадия	Лист	Листов
Разраб.	Астахов С.В.							11	11
Руков.	Григорьев Ю.А.								
Н. Контр.								МГТУ им. Н.Э. Баумана Группа ИУ6-22М	