

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_о_заказчике*, данные_о_заказе*,
срок_выполнения_заказа, фактическая_дата_выполнения, [запчасть*], [ремонт*],
стоимость, id_мастера}}
```

Мастер:

```
{index, doc_type, id, body: {сведения_о_мастере*, [отзыв_о_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_о_заказчике, стоимость_заказа) -
Выполнил(срок_выполнения_заказа, фактическая_дата_выполнения) - Мастер(id_мастера, сведения_о_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		
Взам. инб №		
Подп. и дата		
Инб № подл.		

Разработка макета аналитической системы (Вариант №1)

Изм. Кол.уч Лист № док Подп. Дата

Разраб. Астахов С.В.

Руков. Григорьев Ю.А.

Н. Контр.

Название темы курсового проекта, задание и описание
варианта

Стадия

Лист

Листов

1

11

МГТУ им. Н.Э. Баумана
Группа ИУ6-22М

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"
      }
    }
  }
}
```

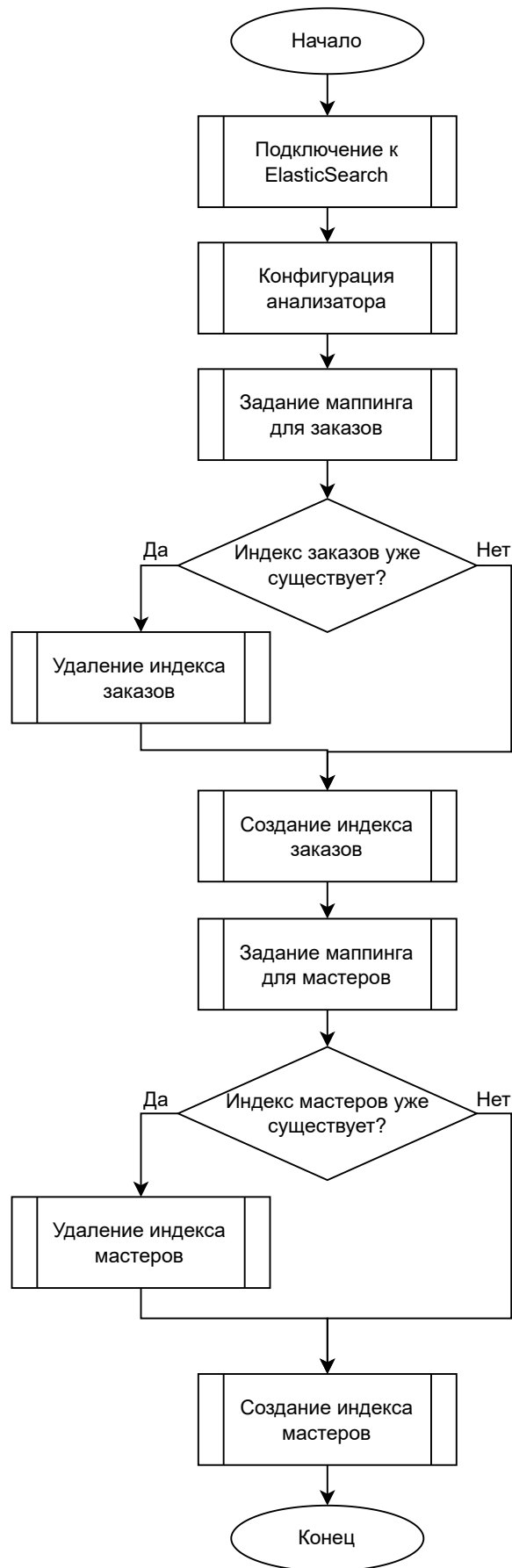
Конфигурация анализатора:

```
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. инв №			
Подп. и дата			
Инв № подл.			

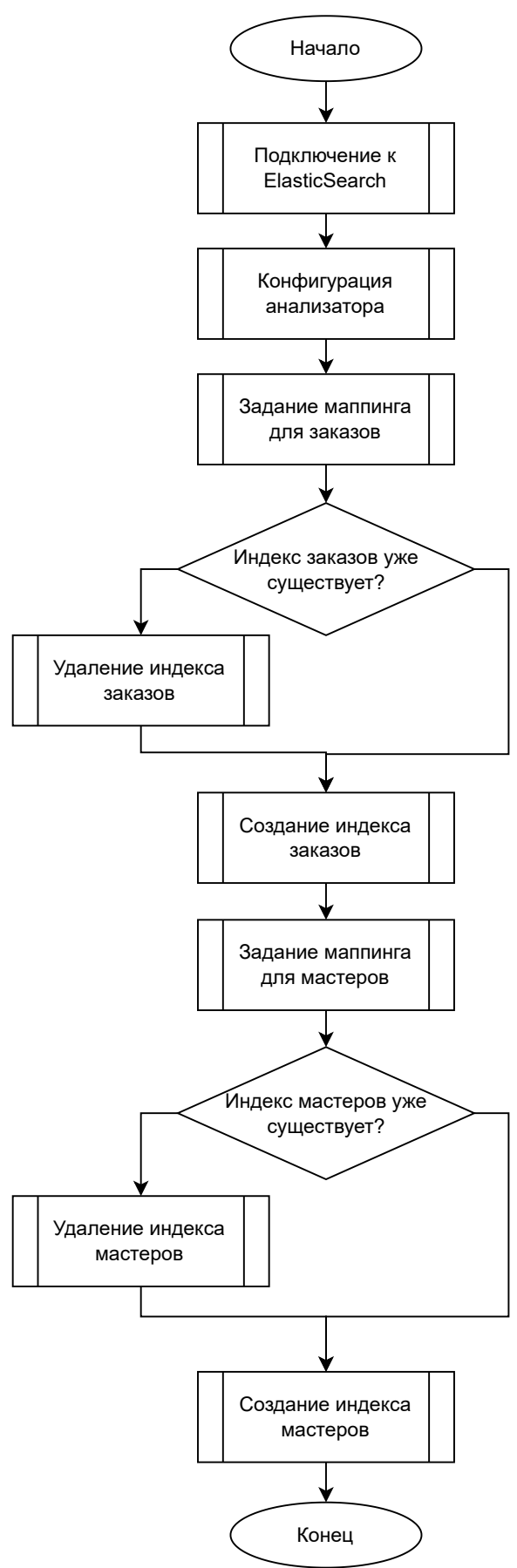
						Разработка макета аналитической системы (Вариант №1)			
Изм.	Кол.уч	Лист	№ док	Подп.	Дата				
Разраб.		Астахов С.В.				Elasticsearch Описание анализатора и маппинга	Стадия	Лист	Листов
Руков.		Григорьев Ю.А.						2	11
							МГТУ им. Н.Э. Баумана Группа ИУ6-22М		
Н. Контр.									

Схема алгоритма программы индексации



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

						Разработка макета аналитической системы (Вариант №1)			
Изм.	Кол.уч	Лист	№ док	Подп.	Дата				
Разраб.		Астахов С.В.				Elasticsearch Схема алгоритма программы индексации документов	Стадия	Лист	Листов
Рукоб.		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			
Взам. инв. №			
Подп. и дата			
Инв. № подл.	Разраб.	Астахов С.В.	
	Руков.	Григорьев Ю.А.	
	Н. Контр.		

						Разработка макета аналитической системы (Вариант №1)			
Изм.	Кол.уч	Лист	№ док	Подп.	Дата				
Разраб.		Астахов С.В.				Elasticsearch Текст и результаты выполнения запросов	Стадия	Лист	Листов
Руков.		Григорьев Ю.А.						4	11
							МГТУ им. Н.Э. Баумана Группа ИУ6-22М		
Н. Контр.									

Текст и результаты выполнения запросов

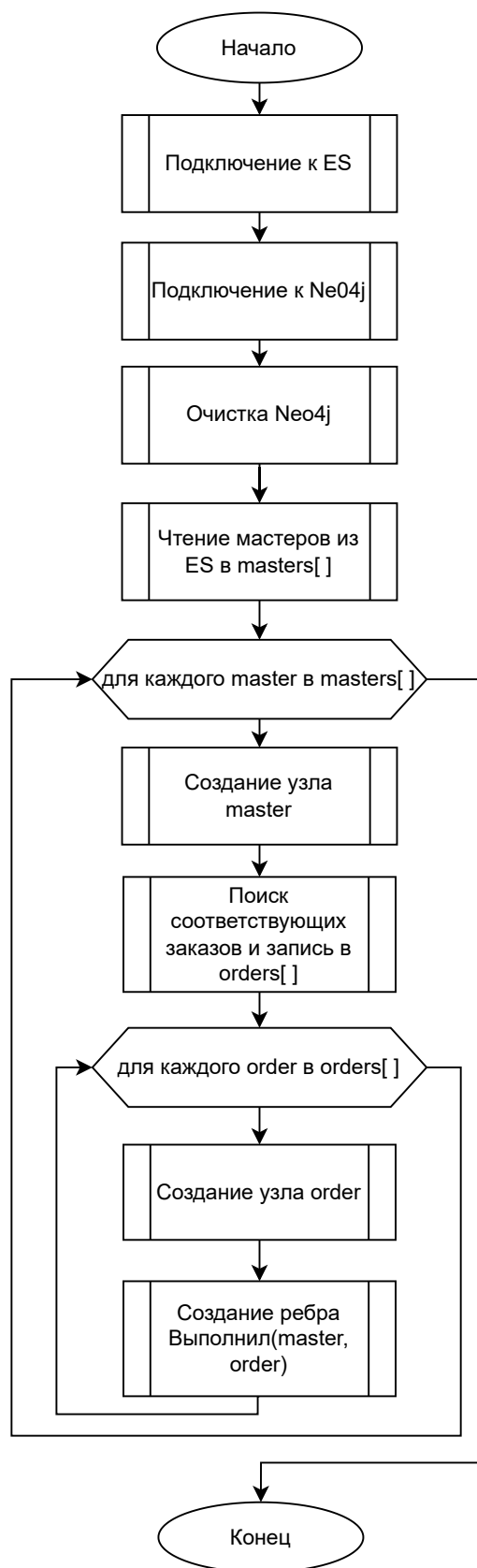
Результат выполнения запроса №2:

```
{'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.	Разраб.	Астахов С.В.	
	Руков.	Григорьев Ю.А.	
	Н. Контр.		

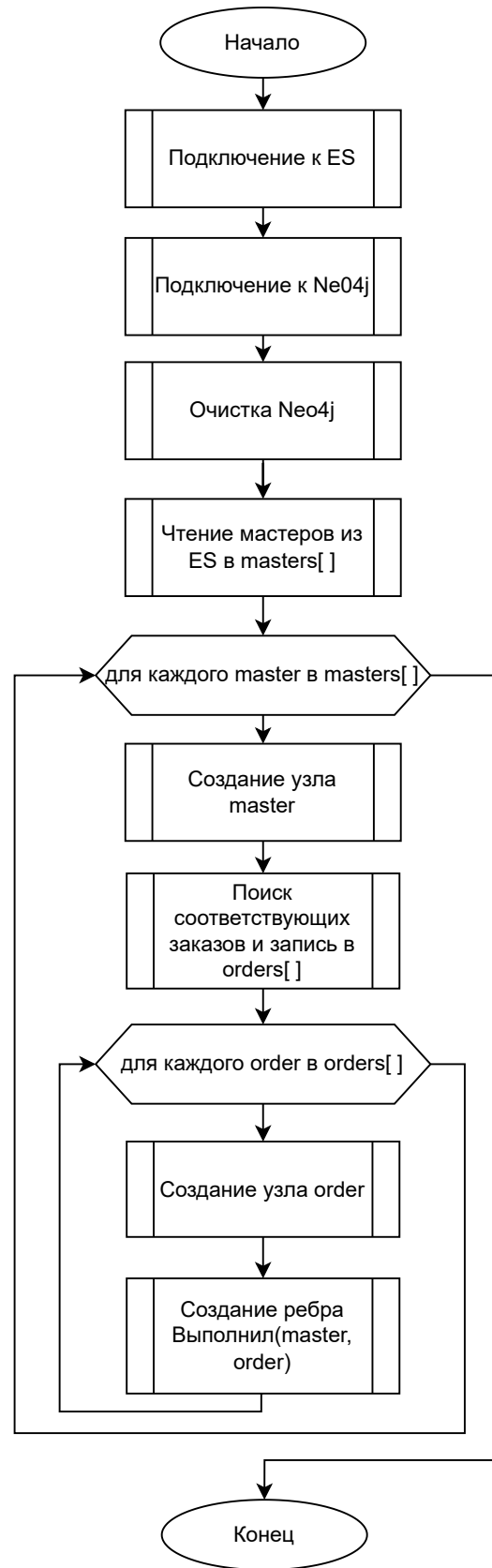
						Разработка макета аналитической системы (Вариант №1)			
Изм.	Кол.уч	Лист	№ док	Подп.	Дата				
Разраб.		Астахов С.В.				Elasticsearch Текст и результаты выполнения запросов	Стадия	Лист	Листов
Руков.		Григорьев Ю.А.						5	11
							МГТУ им. Н.Э. Баумана Группа ИУ6-22М		
Н. Контр.									

Схема алгоритма создания и заполнения графовой базы данных



Инд № подл.	Подп. и дата	Взам. инд №	Создано		

						Разработка макета аналитической системы (Вариант №1)			
Изм.	Кол.уч	Лист	№ док	Подп.	Дата				
Разраб.		Астахов С.В.				Neo4j Схема алгоритма создания и заполнения графовой базы данных	Стадия	Лист	Листов
Руков.		Григорьев Ю.А.						6	11
							МГТУ им. Н.Э. Баумана Группа ИУ6-22М		
Н. Контр.									



Текст запроса на языке Cypher и результаты выполнения запроса

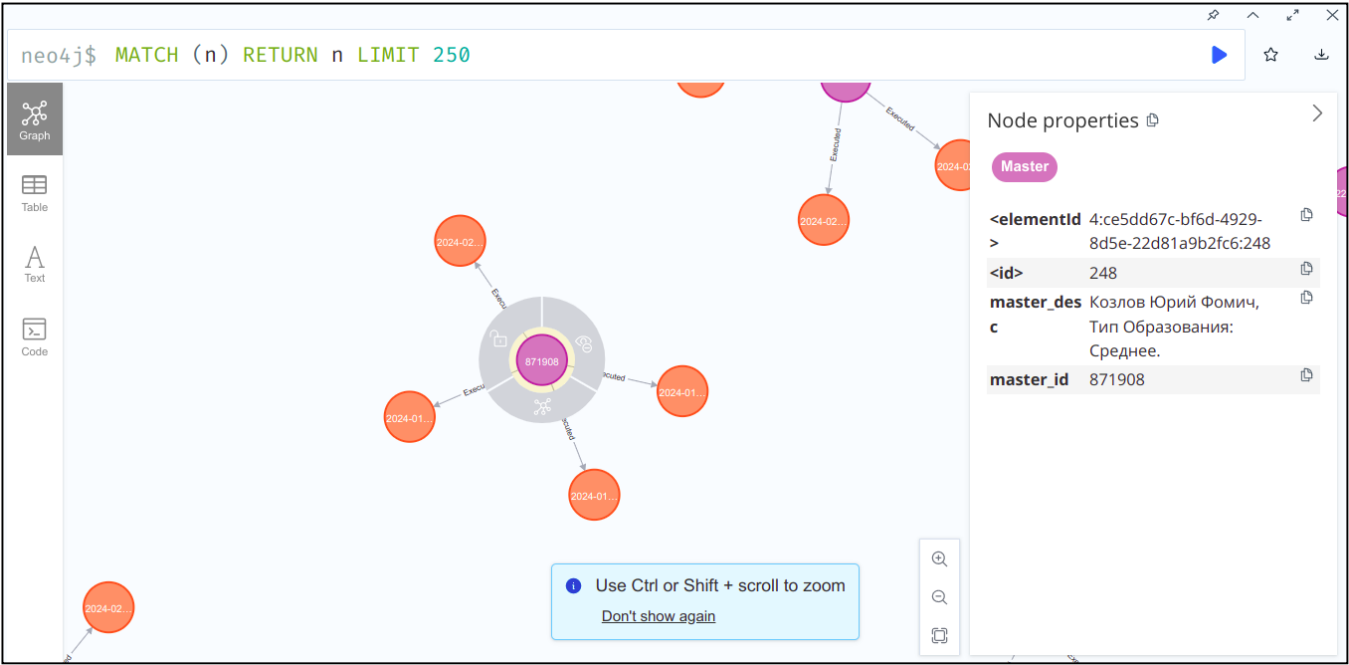
Текст запроса на языке Cypher:

```
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:



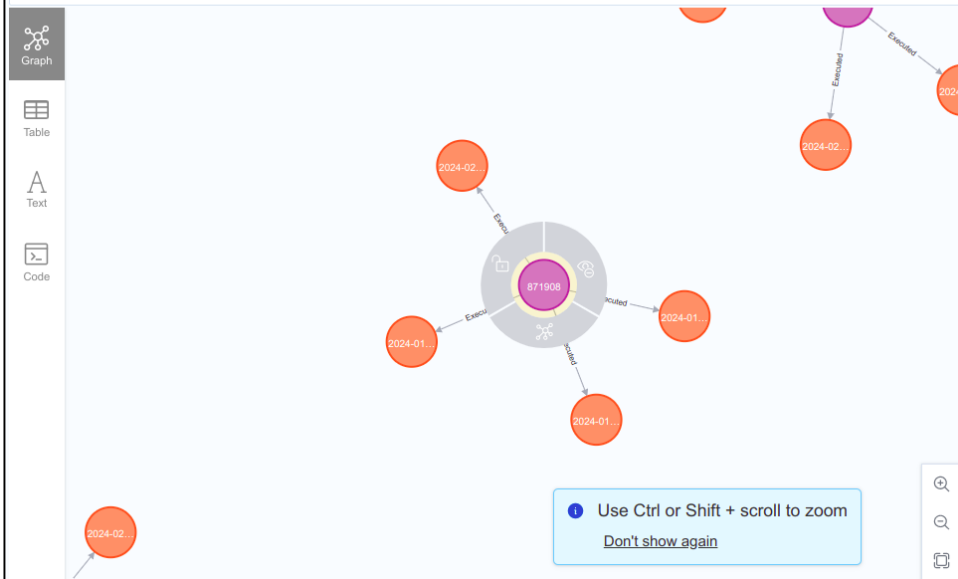
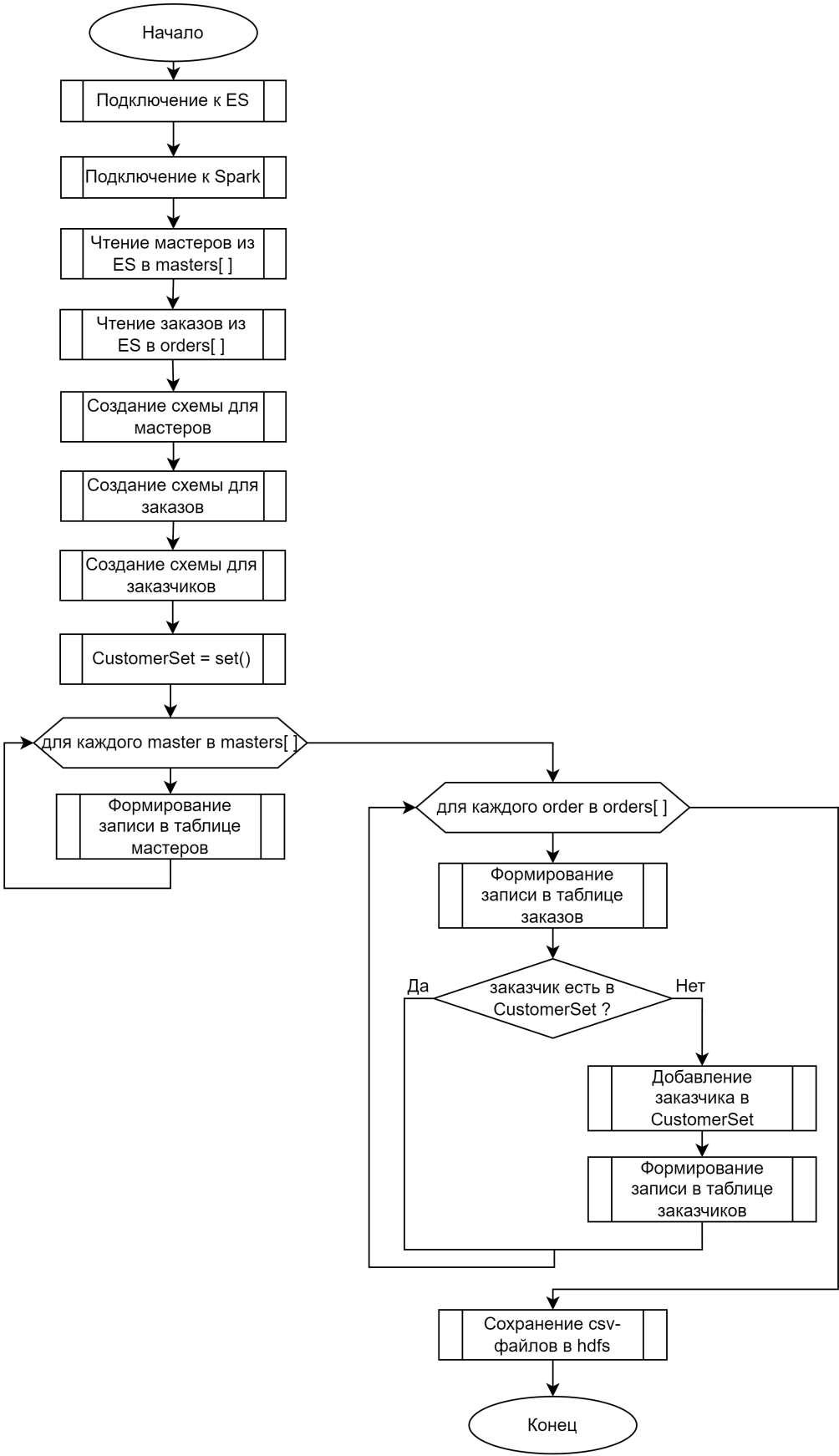
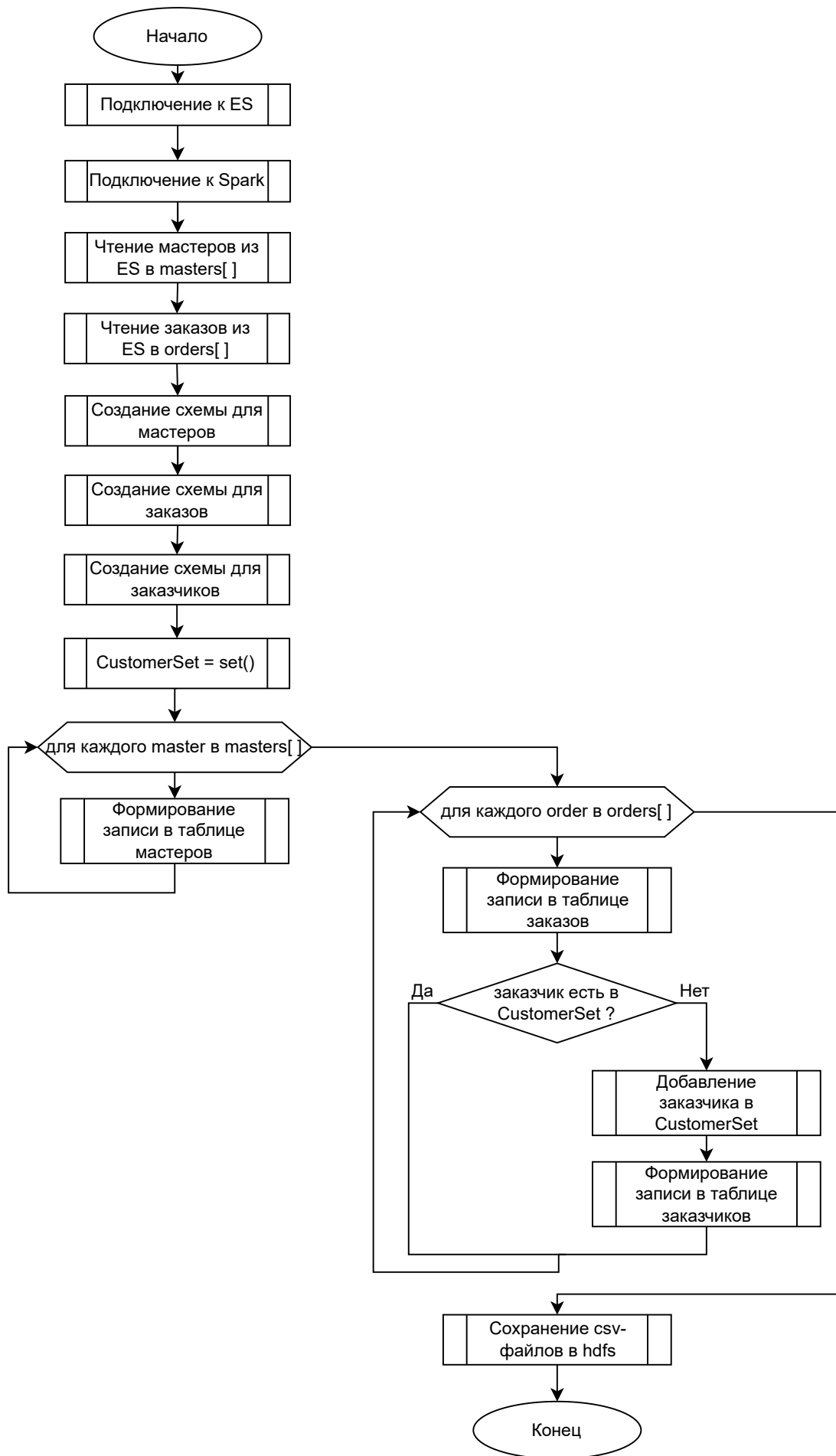
Согласовано						<div><div>neo4j\$ MATCH (n) RETURN n LIMIT 250</div><div><div><div>Graph</div><div>Table</div><div>Text</div><div>Code</div></div><div><div>Node properties</div><div><div>Master</div><div><elementId> 4:ce5dd67c-bf6d-4929-8d5e-22d81a9b2fc6:248</div><div><id> 248</div><div>master_des Козлов Юрий Фомич,</div><div>c Тип Образования: Среднее.</div><div>master_id 871908</div></div></div><div>Use Ctrl or Shift + scroll to zoom</div><div>Don't show again</div></div></div>									
Взам. инв №						Разработка макета аналитической системы (Вариант №1)									
Подп. и дата															
	Изм.	Кол.уч	Лист	№ док	Подп.										
Инв № подл.	Разраб.	Астахов С.В.				Neo4j Текст запроса на языке Сурpher 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. инв №			
Подп. и дата			
Инв № подл.	Разраб.	Астахов С.В.	
	Руков.	Григорьев Ю.А.	
	Н. Контр.		

						Разработка макета аналитической системы (Вариант №1)			
Изм.	Кол.уч	Лист	№ док	Подп.	Дата				
Разраб.		Астахов С.В.				Spark Схема алгоритма создания CSV-файлов	Стадия	Лист	Листов
Руков.		Григорьев Ю.А.						8	11
							МГТУ им. Н.Э. Баумана Группа ИУ6-22М		
Н. Контр.									



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:

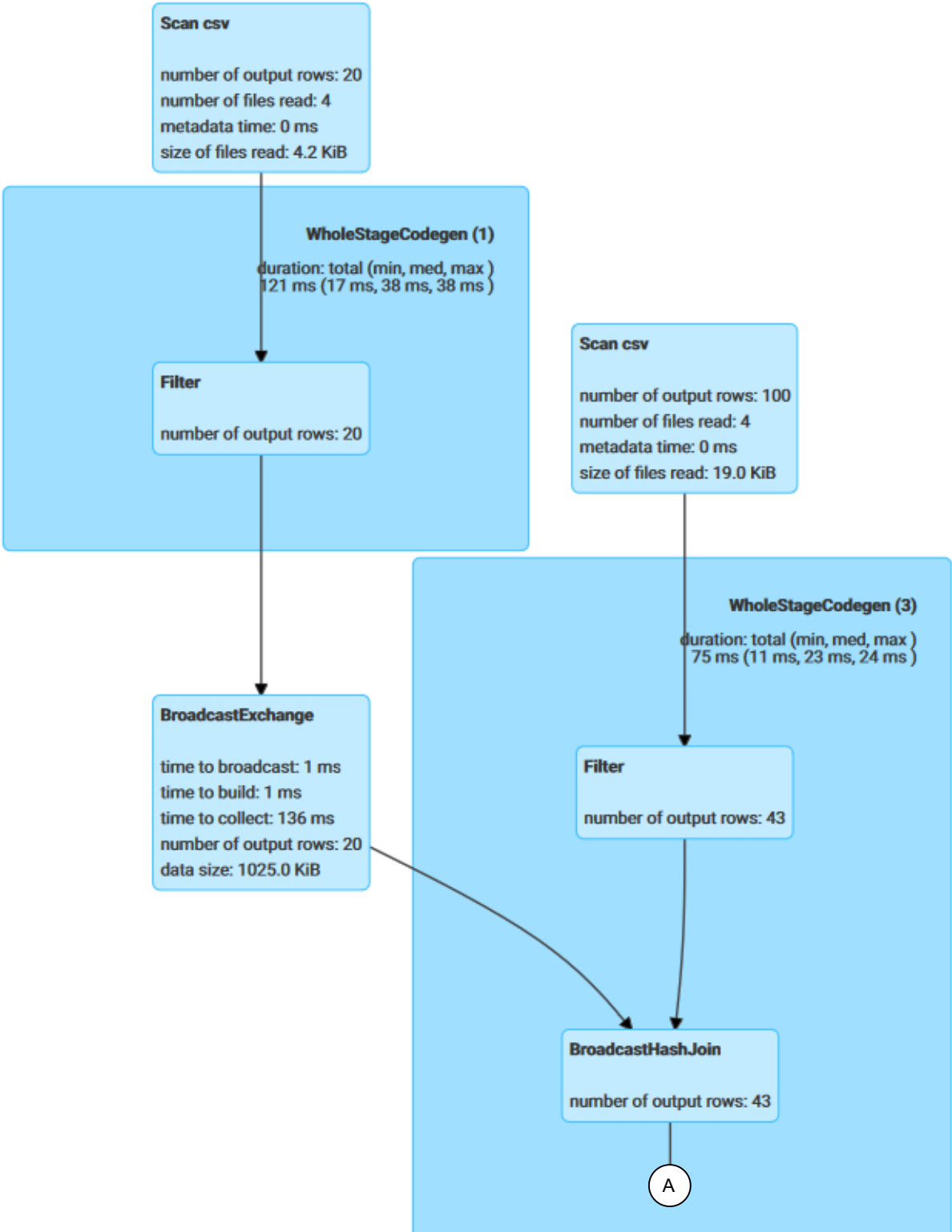
master_id	master_desc	order_id	order_date	due_date	fact_compl_date	customer_id	customer_desc
257458	шестакова на...	131919	2024-01-06	2024-04-10	2024-04-27	227431	имя: маргари...
814777	шарова жанна...	141550	2024-01-10	2024-04-03	2024-04-27	683623	имя: лора бо...
871908	козлов юрий ...	993858	2024-01-16	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	максимильян ...	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.	Кол.уч	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

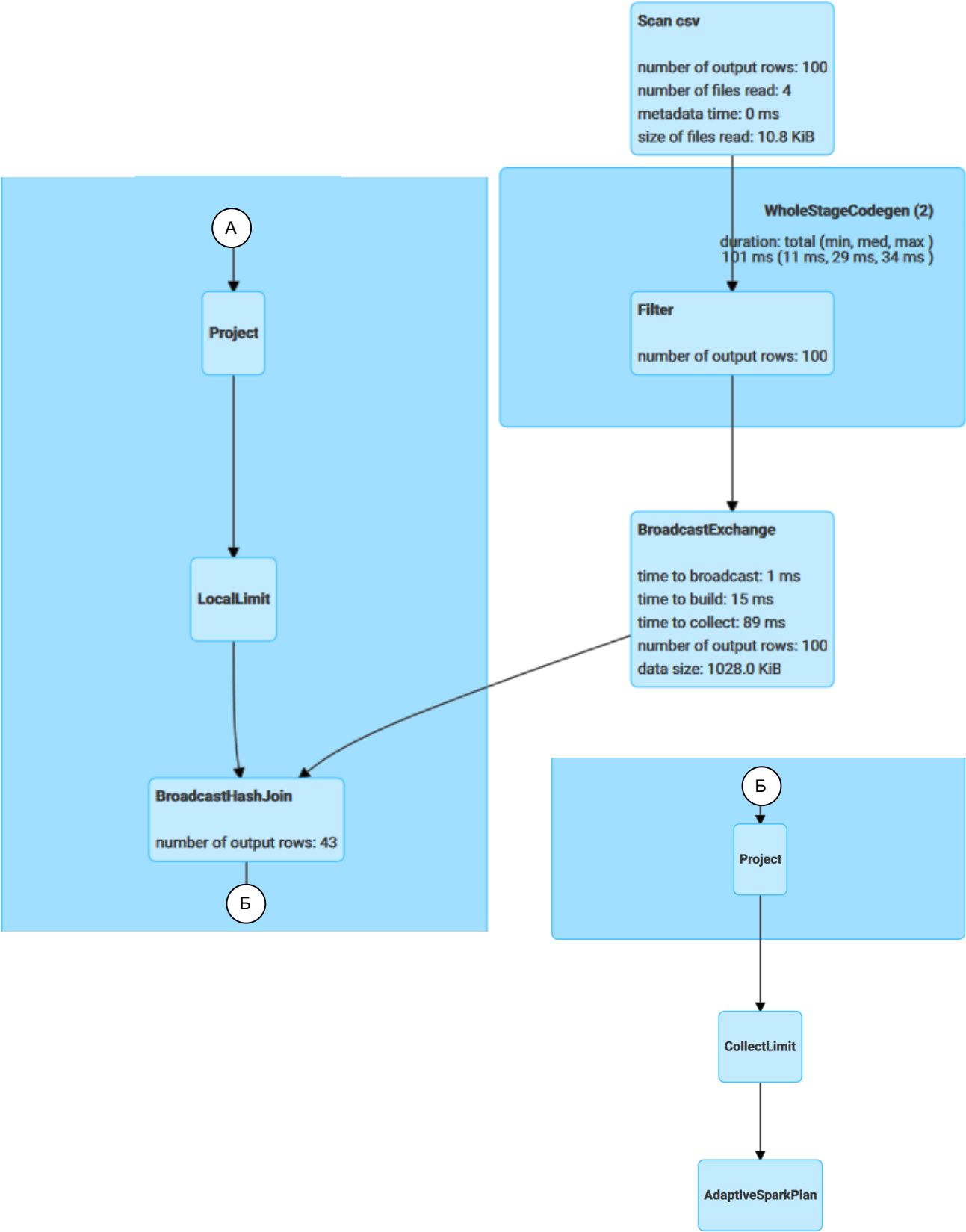


Согласовано			

Взам. инв №	
Подп. и дата	
Инв № подл.	

						Разработка макета аналитической системы (Вариант №1)			
Изм.	Кол.уч	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			
Взам. инв №			
Подп. и дата			
Инв № подл.	Разраб.	Астахов С.В.	
	Руков.	Григорьев Ю.А.	
	Н. Контр.		

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